# Probing chromosome structure using multidimensional scaling of DNA contact matrices

Anthony David Riley

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Department of Statistics

August 2014

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# Acknowledgements

I would like to give my principal thanks to the University of Leeds and EPSRC, for providing me with the opportunity to further my study of statistics. I am also grateful for the funding they have provided me with; the facilities to work in and the opportunity they have given me to visit conferences to share my findings and receive the findings of others.

I would like to thank my supervisors Professor Walter Gilks, Professor Kanti Mardia and Professor John Kent, for their advice, ideas and guidance. These have proved invaluable in for my research. The research skills they have passed down to me will continue to prove invaluable in whatever my future career choice should be. I would also like to thank the other statistical staff in the university, who have answered my simple statistical queries.

I would like to give my thanks to my family and friends, for the patience and support they have provided over the time I have spent on my research.

Finally I am also grateful to my fellow Ph.D. students, who have made my office an interesting, entertaining and fun place to study.

# Abstract

Chromosome conformation capture technology has provided a route to studying genome structure through DNA-DNA contact-counts. An iteration of chromosome conformation capture technology is Hi-C, which provides genome wide two dimensional contact-count data. The contact-count data from Hi-C can be viewed as a proxy for distance and using some transform function can be transformed into estimated distances. These estimated distances can be fitted into Euclidean space using the statistical tools of multidimensional scaling to give estimated chromosome or genome configurations.

The first part of this thesis takes the Hi-C contact-count data for Chromosome 14, transforms it into estimated distances which are fitted into Euclidean space to give an estimated chromosome configuration. Steps are also taken to pre-process the genome contact-count matrix to refine the information held within it. The pre-processed genome contact-count matrix is transformed into estimated distances, which are fitted into Euclidean space to give an estimated genome configuration. The estimated chromosome and genome configurations are investigated, to find if known features of these structures are captured through fitting the Hi-C data.

The second part of this thesis simulates contact-count data from simple configurations. Using the inverse of the transform functions the distances between points in a configuration can be transformed into mean contact-counts. The mean contact-counts are perturbed using a suitable distribution function to provide perturbed contact-counts, which are transformed into perturbed distances. The perturbed distances can be fitted into Euclidean space to give a fitted configurations. The properties of the fitted configurations are investigated and compared with the original configurations, and the properties of the perturbed distances are also investigated. Then steps are taken to improve the fitted configurations using information from the properties of the perturbed distances, with the successful techniques applied to estimating the chromosome configuration.

# Contents

# List of figures

# List of tables

# Chapter 1

# Introduction

The genome is the term used to describe the genetic material of an organism and the nuclear machinery used to maintain and process it. The largest components of the genome are chromosomes, these are large molecules of deoxyribonucleic acid (DNA) and are carriers of genetic information. DNA is a filament, constructed from a chain of millions of base pairs, each pair consisting of adenine with thymine or guanine with cytosine (GC). In Eukaryotes chromosomes are housed in the nucleus and the human nucleus consists of twenty two chromosome pairs and XX for females or XY for males. When unwound the chromosome filament for a human totals approximately two meters in length, approximately 200,000 times the size of the diameter of an average mammalian cell nucleus (de Wit and de Laat, 2012).

The nucleus also houses the nuclear machinery used to maintain and process the chromosomes such as transcription factories (Verschure et al., 1999) which copy DNA into RNA, Cohesin and CTCF proteins which maintain chromosome structure (Hadjur and Sofueva, 2012) and the nuclear lamina (Akhtar and Gasser, 2007; Guelen et al., 2008) to which chromosomes anchor. The quantity of DNA housed in the nucleus and the requirement for genomic machinery to access specific parts of it, means a high level of non-random structure is required for the genome to function and function efficiently.

Elucidating genome structure will further our understanding of genome function.

## 1.1 Interphase genome

The interphase state is when the cell is performing its assigned function and the genome is in a less condensed state allowing the nuclear machinery to freely operate on the chromosomes. Chromosomes in interphase cannot be easily distinguished through a microscope and their structure is largely unknown. What is known is that the DNA molecules are arranged on multiple levels to form chromosomes. The DNA is first wound around histone proteins to produce a 10nm cord, which takes a "beads on a string" appearance. The cord is further wound to produce a 30nm cord. The 10nm cord is called euchromatin and is characterised by gene-rich and open DNA easier for transcription, while the 30nm cord is heterochromatin which is gene-poor and more condensed taking a structural role. Both euchromatin and heterochromatin are examples of chromatin. Research indicates that on the next level of arrangement, chromatin clusters together to form globules (Baù et al., 2011; Goetze et al., 2007; Sanyal et al., 2011) taking another "beads on a string" appearance. The cores of the globules consist of gene-rich active DNA, while gene-poor inactive DNA is found more peripherally. Further arrangement of DNA appears to be driven by the distribution of euchromatin and heterochromatin. Shopland et al. (2006) observed gene-rich chromatin segments partitioned by gene deserts (gene-poor chromatin segments), then the segments making several arrangements, where one feature was clustering of gene-rich segments and clustering of gene deserts. Lieberman-Aiden et al. (2009) and Sanyal et al. (2011) note the genome being partitioned into gene-rich and gene-poor compartments. Compartments could be an additional level of arrangement and formed by clusters of euchromatin and clusters of heterochromatin, with gene-rich DNA located at the interior for easy access to the nuclear machinery and gene-poor DNA located peripherally to provide structural support (Guelen et al.,

2008) and to avoid inhibiting the nuclear machinery. DNA-DNA contacts made between genomically distant (along the DNA filament) segments (Lieberman-Aiden et al., 2009; Simonis et al., 2006) provide evidence for chromatin clustering.

The chromosomes are arranged into chromosome territories (CT's) (Cremer et al., 1993; Cremer and Cremer, 2010; Heard and Bickmore, 2007), which make few interchromosomal interactions (Lieberman-Aiden et al., 2009). CT arrangement in the nucleus is measured on radial distance from the origin (centre) of the nucleus. One proposal is that CT's are radially arranged according to gene-content (Boyle et al., 2001; Tanabe et al., 2002), where gene-rich chromosomes localize at the nuclear interior and gene-poor chromosomes localize at the nuclear periphery. Another idea proposes CT's are radially positioned according to chromosome size (Bolzer et al., 2005), with chromosome size increasing as radial distance increases. Another idea blends the two ideas and proposes CT's are radially positioned according to the ratio of gene-density to chromosome size (Heride et al., 2010). If CT's radial positioning is driven by gene-content then this appears to follow a trend of gene-rich material colocalizing in the interior of the structures (globule, chromosome or nucleus) and gene-poor material localizing on the periphery. Genome structure can be imagined as balls of string sitting in a basket, where each ball of string is made from a single fibre wound up, and inside some of the balls the string has unwound leaving clumps of free fibre.

## 1.2 Probing genome structure

In the latter half of the twentieth century microscopy was used to study genome structure. The principle method of microscopy was fluorescence in situ hybridization (FISH). The FISH method attached fluorescent probes to specific DNA (or RNA) sequences and recorded the probes position in the genome when observed through a microscope. FISH helped decipher chromosome territories arrangement, the physical properties of

chromatin and investigated the folding patterns of gene-rich and gene-poor chromosome regions. The FISH method is limited by the amount of fluorescent probes which can be attached and ultimately microscope resolution. In 2002 Dekker et al. developed the technique of chromosome conformation capture (3C) which opened the way to studying the genome using DNA-DNA contacts, where a contact signifies two DNA segments are spatially close approximately 10nm-100nm (Dekker et al., 2013). The advent of 3C and associated technologies 4C, 5C and Hi-C has allowed the study of DNA-DNA contacts, providing a new avenue to study genome structure and eventually reconstructing a 3D low-resolution path of chromatin through the nucleus.

### 1.2.1 Chromosome conformation capture

Invented by Dekker et al. (2002), chromosome conformation capture (3C) provides one-to-one contact frequency data between selected sites on the genome. The method is briefly described below, with an illustration in Figure 1.2.

1. Formaldehyde is used to cross-link (fix) spatially close DNA segments to each other.
2. Cross-linked DNA is sheared from non-cross-linked DNA using a restriction enzyme, such as HindIII.
3. Single ends of the cross-linked segments are ligated, producing new DNA segment resembling a loop.
4. Cross-linking is reversed leaving ligated DNA segments, with each segment is the union of two previously spatially close segments.
5. Polymerase chain reaction (PCR) of selected ligation junctions is used to semi quantitatively assess the frequency with which sites of interest make contact.

Repeating 3C can provide a one dimensional data set of DNA-DNA contact frequency between a site of interest on a chromosome. DNA-DNA contact frequency (counts) measures how many times contacts are detected between two sites, using a 3C based method.

### 1.2.2 Chromosome conformation capture-on-chip

Invented by Simonis et al. (2006), chromosome conformation capture-on-chip (4C) provides a one dimensional (one-to-all) count data between a site of interest and all the sites on the genome. The method is briefly described below, with illustration in Figure 1.3.

1.- 4. As in 3C process.

5. Second round of shearing and ligation is applied to the ligated DNA segments, shortening then linking the freshly sheared ends to produce DNA circles.

6. Inverse PCR using primers specific to the site of interest, amplify all the segments contacting the site of interest.

7. The pool of amplified DNA-DNA contacts data is then counted using large scale sequencing.

### 1.2.3 Chromosome conformation capture carbon copy

Invented by Dostie et al. (2006), chromosome conformation capture carbon copy (5C) provides two dimensional (many-to-many) data between sites of interest on the genome. The method is briefly described below, with illustration in Figure 1.4.

1.- 4. As in 3C process.

5. The new DNA segments are mixed with oligonucleotides, which partially overlap the different restriction sites of interest.

6. Oligonucleotides corresponding to the new DNA segments are juxtaposed and ligated to them to produce new ligation products.

7. The new ligation products are simultaneously amplified and counted using large scale sequencing, to produce symmetric matrix of counts between the sites of interest.

### 1.2.4 Hi-C

Invented by (Lieberman-Aiden et al., 2009), Hi-C provides two dimensional (all-to-all) count data between all the sites on the genome. The method is briefly described below, with illustration in Figure 1.5.

1.- 2. As in 3C process.

3. Restriction ends are filled using biotin-labelled nucleotides then blunt end ligation is performed.

4. DNA segments are then purified and sheared, leaving the ligation junctions which have been tagged with the biotin pull down.

5. Ligation junctions are mapped back to the genome using large scale sequencing, to produce a symmetric matrix of counts between all the sites on the genome.

Chromosome conformation capture based technologies can be abstractly imagined to help understand them. Imagine the genome as the string in the basket analogy described earlier, but the string is now multicoloured where colours correspond to different regions of the chromosome.

| | Mb 1 | Mb 2 | Mb 3 | Mb 4 | Mb 5 | Mb 6 | Mb 7 | Mb 8 | Mb 9 | Mb 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Mb 1 | 2767 | 527 | 113 | 88 | 123 | 190 | 166 | 109 | 118 | 117 |
| Mb 2 | 527 | 3826 | 440 | 239 | 261 | 183 | 63 | 43 | 23 | 54 |
| Mb 3 | 113 | 440 | 3522 | 948 | 341 | 156 | 44 | 24 | 25 | 44 |
| Mb 4 | 88 | 239 | 948 | 5156 | 876 | 139 | 35 | 21 | 19 | 30 |
| Mb 5 | 123 | 261 | 341 | 876 | 5703 | 492 | 76 | 42 | 27 | 71 |
| Mb 6 | 190 | 183 | 156 | 139 | 492 | 3854 | 372 | 173 | 132 | 192 |
| Mb 7 | 166 | 63 | 44 | 35 | 76 | 372 | 2684 | 501 | 342 | 231 |
| Mb 8 | 109 | 43 | 24 | 21 | 42 | 173 | 501 | 2311 | 530 | 259 |
| Mb 9 | 118 | 23 | 25 | 19 | 27 | 132 | 342 | 530 | 2096 | 385 |
| Mb 10 | 117 | 54 | 44 | 30 | 71 | 192 | 231 | 259 | 385 | 2766 |

Table 1.1: Sample of the Hi-C count matrix (Lieberman-Aiden et al., 2009), recording the DNA-DNA contacts made between megabase intervals 1 to 10 in Chromosome 14, from a karyotypically normal lymphoblastoid cell line.

1. Pour hot toffee through the string and basket, residual blobs of toffee will stick close segments of string together.

2. Cut the string up, leaving lots of pieces of string with some of the pieces attached to each other with the blob of toffee.

3. Tie the attached pairs together at one end making loops of string.

4. Chew the toffee off the loops to leave new string segments, with each segment half from a previously close string segment.

5. Count up the different coloured segments and record, i.e. how many red and green segments are attached or red and red segments are attached to each other.

### 1.2.5 Count data to genome structure

Two dimensional chromosome contact data provided by 5C and Hi-C provides a promising route to elucidate genome structure. Table 1.1 gives a sample of a Hi-C count matrix, the larger the counts the spatially closer two megabase intervals should be. The

Hi-C method in particular provides low resolution ($\approx$ 1Mb) data which can be used to reconstruct the average path of megabase intervals through the genome. Although the two dimensional count data presents several issues for analysis. The first issue is that (illustrated in Figure 1.1), human chromosomes are diploid which means chromosomes come in identical pairs (in Figure 1.1 this is illustrated by chromosome A or chromosome B). This provides two types of count, intrachromosomal (within the chromosome) counts are generated from intrachromosomal contacts made on intervals within chromosome A and B; or interchromosomal (between the chromosomes) counts made between the same interval on chromosome A and B. Second, the genome is labile so the count data represents a cell average map of the genomes in differing structures. Finally, the Hi-C data contains several experimental biases such as the distance between restriction sites, the mappabilty of the sites, trimmed ligation junctions, Guanine and Cytosine (GC) content and the distance between restriction sites (Yaffe and Tanay, 2011).

Two approaches have been taken to elucidate genome structure (Dekker et al., 2013). One approach involves modelling the chromatin as a polymer and using random walk mathematics to explain contact frequency (counts). Another approach uses the counts to estimate a spatial distance, treats the problem as an optimization problem and solves it to find a three dimensional estimated chromosome configuration which minimizes a score function.

In addition to genome structure, research has uncovered and corrected for biases in Hi-C data and uncovered topological properties of the genome. Yaffe and Tanay (2011) identifies several biases in the Hi-C data such as distance between restriction sites, GC content of trimmed ligation junctions and sequence uniqueness, corrects for the biases using a probabilistic model and identifies topological features of the genome such as interchromosomal contacts between GC-content domains and intrachromosomal contacts around transcription start sites. Lieberman-Aiden et al. (2009) used principle component analysis on a normalized Hi-C count matrix, to identify that the genome is partitioned

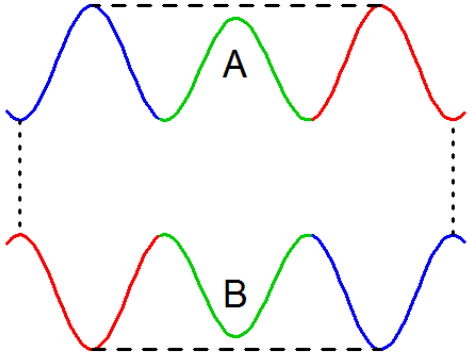



Figure 1.1: Illustration of potential cross contacts made between the identical chromosomes in a homologous pair. The long dashed line ----, signifies intrachromosomal contacts made between different megabase intervals on the chromosome A or chromosome B. The short dashed line ....., signifies the interchromosomal contacts made between different megabase intervals of the chromosome pair. These interchromosomal contacts are confused with intrachromosomal contacts as chromosomes A or B are identical. The coloured lines —,— and — denote different megabase intervals on the chromosome, which are identical to the megabase interval (of the same colour) on the other chromosome.

into two compartments, compartment A consisting of gene-rich like euchromatin and compartment B consisting of gene-poor and dense heterochromatin. Other approaches have been used to try to elucidate genome structure such as complex DNA-DNA network models (Kruse et al., 2013).

### 1.2.6 Polymer models

Prior to the development of 3C technologies, polymer models were used to try to explain chromatin structure. Polymer models try to predict the chromosome structure to explain the data (top-down). Polymer modelling can be used to interpret the distribution of loop sizes, formation of CT's, shape of CT's and physical properties of euchromatin

and heterochromatin. Sachs et al. (1995) used FISH data to interpret the relationship between genomic and Euclidean distances, observing giant chromatin loops ($\approx$ 3Mb long) and that chromatin conforms to the sample path of a loose (less compact) random walk at low genomic distance ($\leq$ 2Mb) but conforms to the sample path of a tight (more compact) random walk at large genomic distances ($\geq$ 2Mb). The 3C technologies have shown chromatin can form loops of differing lengths (Lieberman-Aiden et al., 2009; Simonis et al., 2006). Two models try to explain this. The dynamic random loops model (Mateos-Langerak et al., 2009) and the fractal globule model (Lieberman-Aiden et al., 2009). The dynamic random loop model recognises that Euclidean distances cannot continue to increase with increasing genomic distance, so it attempts to model Euclidean distance at small genomic distance and Euclidean distance becomes independent of genomic distance at large genomic distance. The fractal globule model assumes that chromatin condenses into a knot free globule, where the chromatin condenses into knot free structures (resembling a beads on a string configuration) which repeatedly fold into further knot free structures, providing structure while allowing segments of chromatin easily to unfold when required. Lieberman-Aiden et al. (2009) used simulations to show that the relationship between contact probability and genomic distance, is better reflected in the fractal globule model than the equilibrium globule model, where the equilibrium globule describes a compact and highly knotted structure.

### 1.2.7 Restraint models

Restraint models usually transform two dimensional counted data into restraints, these restraints dictate how close or distant intervals should be placed, according to how the count size has been interpreted. The restraints are inputted into an optimization procedure to find the 3D configuration which best respects the restraints. This has been attempted by numerous groups (Barbieri et al., 2012; Baù et al., 2011; Duan et al., 2010; Hu et al., 2013; Kalhor et al., 2012; Meluzzi and Arya, 2013; Peng et al., 2013; Rousseau et al.,

2011; Trieu and Cheng, 2014; Ben-Elazar et al., 2013) five of which are summarized below.

Duan et al. (2010) combined 4C with massively parallel sequencing to collect count data on the Saccharomyces Cerevisiae (bakers yeast) genome. Duan et al. assumed an observed count has the same estimated distance $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$ as the distance needed to produce the same expected count through polymer packing. Points $\underline{x}_i = (x_{i,1}, x_{i,2}, x_{i,3})$ were fit to minimize the score function

$$\sum_{i<j} (\hat{d}_{i,j} - \tilde{d}_{i,j})^2,$$

where $\hat{d}_{i,j} = (\sum_{k=1}^{3} (\hat{x}_{i,k} - \hat{x}_{j,k})^2)^{\frac{1}{2}}$, with respect to known constraints such as the range of $\hat{d}_{i,j}$ between intrachromosomal or interchromosomal points or between adjacent points. The optimization problem was solved using interior point optimizer (Wächter and Biegler, 2006) and gave a Rabl configuration for the yeast genome, with the centromeres clustering at one end of the nucleus and telomeres clustering at the opposite end.

Peng et al. (2013) corrected experimental sequencing depth bias on the Hi-C data from various human cell types before applying a similar approach to Duan et al. (2010), where experimental sequencing depth is where the distribution of contacts differs at different levels of sequencing depth. The $\tilde{d}_{i,j}$ were constructed from the counts using a linear transform function, where the parameters were determined by the diameter of the fitted points and the chromatin region. Points were fit into 3D space to minimize the score function

$$\sum_{i<j} \frac{(\hat{d}_{i,j} - \tilde{d}_{i,j})^2}{\tilde{d}_{i,j}^2},$$

subject to similar constraints used by Duan et al. (2010). The optimization problem was solved using the automatic pipeline AutoChrom3D.

Baù et al. (2011) combined 5C data with an integrated modelling platform (IMP), to reconstruct the ENm008 domain (500kb) on human chromosome 16, for K562 cells and silenced lymphoblastoid cells. Restraints were constructed by taking the logarithm (base 10) of the observed counts calculating a z-score (using the mean and standard deviation of the logarithm values) and inputting the z-scores into two linear transforms, one for adjacent and one for non-adjacent points. The IMP took the restraints and reconstructed the 3D structure of the domain such that $\hat{d}_{i,j}$ were inversely proportional to observed counts. The reconstruction produced chromatin clusters with active genes and gene promoters internally located and inactive genes and restriction fragments more peripherally located.

Ben-Elazar et al. (2013) used chromosome contact-count data provided by 3C on the yeast genome (Duan et al., 2010), to estimate chromosome structure. The data was first filtered to remove false contact-counts caused by noise in the experimental process. Then natural neighbour interpolation was applied to each chromosome contact-count matrix to produce a matrix of dissimilarities between sites on the chromosome. The dissimilarity matrix was fitted into three dimensional Euclidean space using metric multidimensional scaling (see Section 2.1), to give an initial three dimensional configuration. The initial three dimensional chromosome configuration and dissimilarity matrix were then used by non-metric multidimensional scaling (see Section 2.2), to give a refined three dimensional chromosome configuration in which the interpoint distances better respected the dissimilarities. Ben-Elazar et al. then used the refined three dimensional chromosome configuration and data on gene locations, to find how genes co-localize in three dimensional space.

Hu et al. (2013) used a Bayesian approach to reconstruct the 3D chromosome structure, using a "Bayesian 3D constructor for Hi-C data" (BACH). Observed counts $\mathbf{M} = (m_{i,j})$ were assumed to come from a Poisson distribution with a mean parameter $\mathbf{U} = \mu_{i,j}$,

modelled with the following link function

$$\log(\mu_{i,j}) = \gamma_0 + \gamma_1 \log(d_{i,j}) + \gamma_{\text{enz}} \log(e_i e_j) + \gamma_{\text{gcc}} \log(g_i g_j) + \gamma_{\text{map}} \log(o_i o_j),$$

where $\gamma_0$ determines scale; $\gamma_1$ determines the relationship between $\mu_{i,j}$ and the inter point distances $\mathbf{D} = (d_{i,j})$ from the chromosome configuration $\mathbf{X} = (x_{i,k})$. The $e_i$ represent fragment ends in site $i$; the $g_i$ represents mean GC content in site $i$; the $o_i$ represents mean mappability of fragment ends in site $i$; then $\gamma_{\text{enz}}$, $\gamma_{\text{gcc}}$ and $\gamma_{\text{map}}$ are the respective coefficients stored in the vector $\underline{\gamma} = (\gamma_0, \gamma_1, \gamma_{\text{enz}}, \gamma_{\text{gcc}}, \gamma_{\text{map}})$. The joint likelihood function

$$P(\mathbf{M}|\mathbf{X}, \underline{\gamma}) = \prod_{i<j} \frac{e^{-\mu_{i,j}} \mu_{i,j}^{m_{i,j}}}{m_{i,j}!}$$

combined with non-informative priors gave the joint posterior distribution

$$P(\mathbf{X}, \underline{\gamma}|\mathbf{M}) \propto \prod_{i<j} e^{-\mu_{i,j}} \mu^{m_{i,j}}. \tag{1.1}$$

Samples were drawn from (1.1) by assigning nuisance parameters to $\underline{\gamma}$, then applying sequential importance sampling to generate a fitted configuration $\hat{\mathbf{X}}$ and finally refining $\underline{\gamma}$ and $\hat{\mathbf{X}}$ using a Gibbs sampler with hybrid Monte Carlo and adaptive rejection sampling. Hu et al. (2013) takes into account experimental biases highlighted by Yaffe and Tanay (2011) and the fact that Hi-C data represents a cell average of multiple genome structures.

Varoquaux et al. (2014) also used the Poisson distribution (similar to Hu et al.) to model the observed counts from Hi-C. First the observed counts were corrected for experimental biases, using an iterative correction and eigenvalue decomposition procedure (Imakaev et al., 2012). Then the $m_{i,j}$ were modelled as Poisson random variables with parameter $\mu_{i,j}$. The parameter $\mu_{i,j}$ was found using an inverse count to distance transform function

$$\mu_{i,j} = a_0 d_{i,j}^{\gamma_1}. \tag{1.2}$$

Then using (1.2) a Poisson log-likelihood function was constructed

$$l(\mathbf{X}, a_0, \gamma_1) = \sum_{i<j\leq n} m_{i,j}\gamma_1 \log(d_{i,j}) + m_{i,j}\log(a_0) - a_0 d_{i,j}^{\gamma_1}, \tag{1.3}$$

where $d_{i,j}$ were extracted from the configuration $\mathbf{X} = (x_{i,k})$ using (2.2) with $p = 3$. The log-likelihood function (1.3) was then maximised using two approaches. The first approach inferred $\gamma_1 = -3$ from relationship information provided by Lieberman-Aiden et al. (2009), between the observed count size and genomic distance, and between the genomic distance and Euclidean distance. The $a_0$ parameter was set at $a_0 = 1$ as its role was trivial. Then an estimated chromosome configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$ was found which maximized (1.3). The second approach iteratively estimated the parameter $\gamma_1$ then found $\hat{\mathbf{X}}$ which maximized (1.3), until optimization. Both maximization routines used the IPOPT algorithm (Wächter and Biegler, 2006). The Poisson maximum likelihood approaches were compared with approaches used by other research groups on simulated data, reporting it performing well.

Trieu and Cheng (2014) reconstructs the 3D chromosome configurations directly from counts without transforming into estimated distances. Observed counts were pre-processed to distinguish genuine contacts from spurious contacts (non-contacts), by generating an interaction frequency matrix $\mathbf{F} = (f_{i,j})$

$$f_{i,j} = m_{i,j} \frac{\sum_{i<j} m_{i,j}}{\sum_i m_{i,j} \sum_j m_{i,j}} \tag{1.4}$$

and using a threshold value of 0.66 to differentiate contacts from non-contacts. Points were considered in contact if $\hat{d}_{i,j}$ was less than some threshold distance $d_c$ otherwise they were non-contacts. Taking the contact & non-contact principle into account and additional constraints for adjacent and non-adjacent points, a score function was constructed and optimized using a gradient decent algorithm to recover $\hat{\mathbf{X}}$.

## 1.3 Research undertaken in this thesis

The Hi-C method provides a two dimensional genome wide proximity map which can be used to reconstruct genome structure. Various groups have already attempted to do this. This research project uses Hi-C count data of Karyotypically normal human lymphoblastoid cells at 1Mb resolution from Lieberman-Aiden et al. (2009). Chromosome structure is reconstructed by transforming counts into estimated distances using a count to distance transform functions. Then fitting the estimated distances into three dimensional Euclidean space using the statistical method of multidimensional scaling (MDS). The estimated chromosome configuration is then investigated, by find comparisons in other research and assessing the robustness of the transformation and MDS. Symmetric sub matrices of the chromosome contact matrix are transformed and fitted into Euclidean space to assess how robust the MDS is to large distances and isolate local structure within the chromosome. The global contact matrix recording all the intrachromosomal and interchromosomal contacts is transformed and fitted on multiple resolutions, to assess how robust the fitting is to the sparse information from the interchromosomal contacts, by comparing features in the fitted configurations with known genomic features.

The process of transforming counts into estimated distances and fitting them into three dimensional Euclidean space is investigated using count data generated from known configurations. Since chromosome configurations are unknown there is nothing avalible to compare the estimated chromosome configurations against. Using count data from known configurations allows the fitted configuration to be compared with the original configuration to observe how the noise in the counted data effects the fitted configuration. This method discovered biases which inflate the estimated distances and are detrimental to the fitted configurations.

The effect of the bias on the fitted configurations is investigated, providing a successful

bias correction and dispersion estimation technique for one method of transforming counts into distances and fitting into Euclidean space. The investigation also highlights how a successful bias correction technique is not possible for other methods of transforming and fitting and why properties of the count to distance transform function can be detrimental to the fitted configuration or the estimated chromosome configuration. Techniques of smoothing out noise from the fitted configurations are also discussed.

Properties of metric MDS (one of the MDS methods) are investigated, by dissecting the fitting of small distance matrices into space and investigating how large distortions in distance matrices are managed.

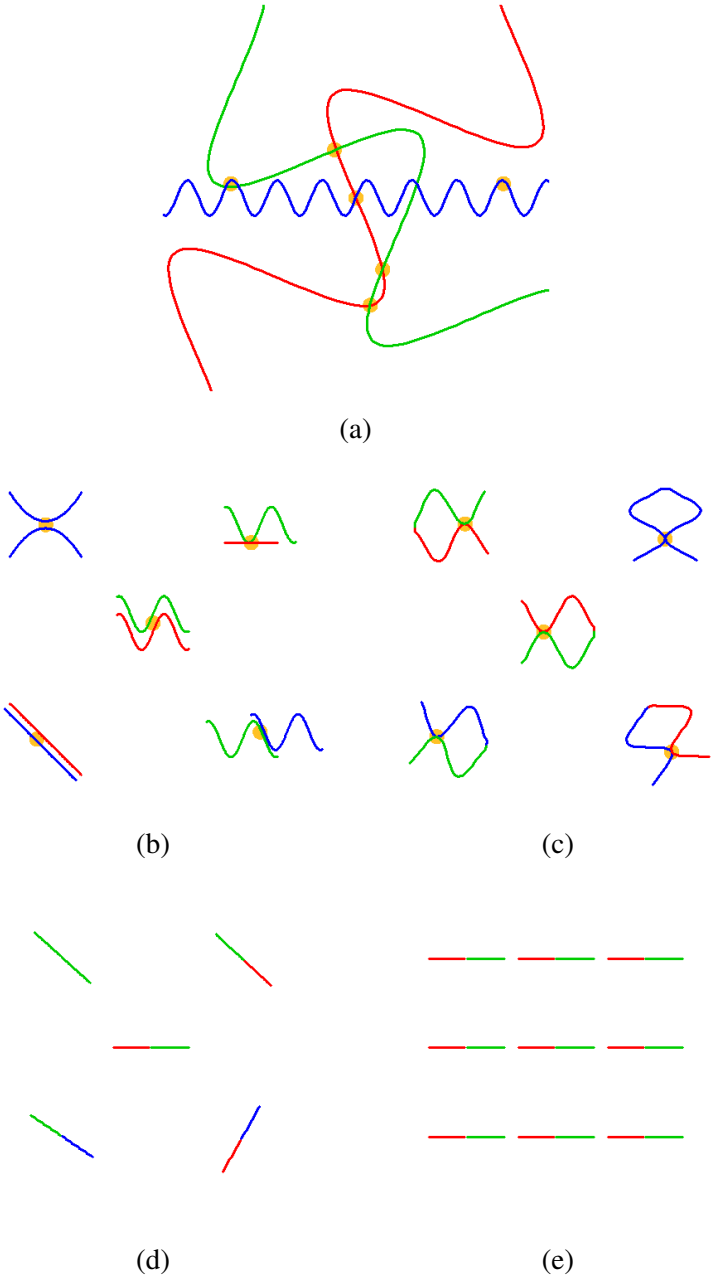
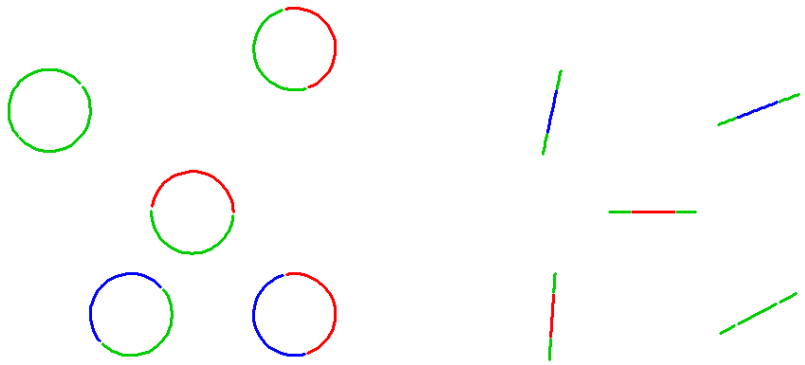
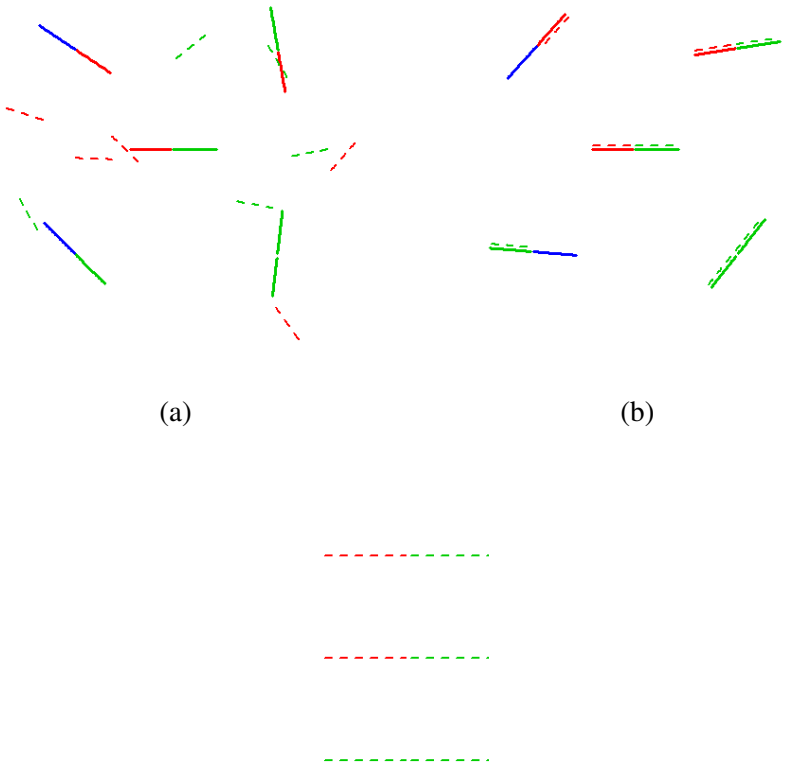



Figure 1.2: Illustration of the 3C process outlined in Section 1.2.1. In Figure 1.2a the formaldehyde fixes spatially close DNA segments. In Figure 1.2b the fixed DNA has been sheared from the fixed DNA. In Figure 1.2c fixed DNA segments are ligated. In Figure 1.2d fixing is reversed leaving the ligated segments. In Figure 1.2e PCR of ligation junctions of interest, allows assessment of frequency sites of interest to make contact. In the figures —, — and — denote DNA originating from different parts of the chromosome (or genome). The formaldehyde is denoted by ●.



(a) (b)

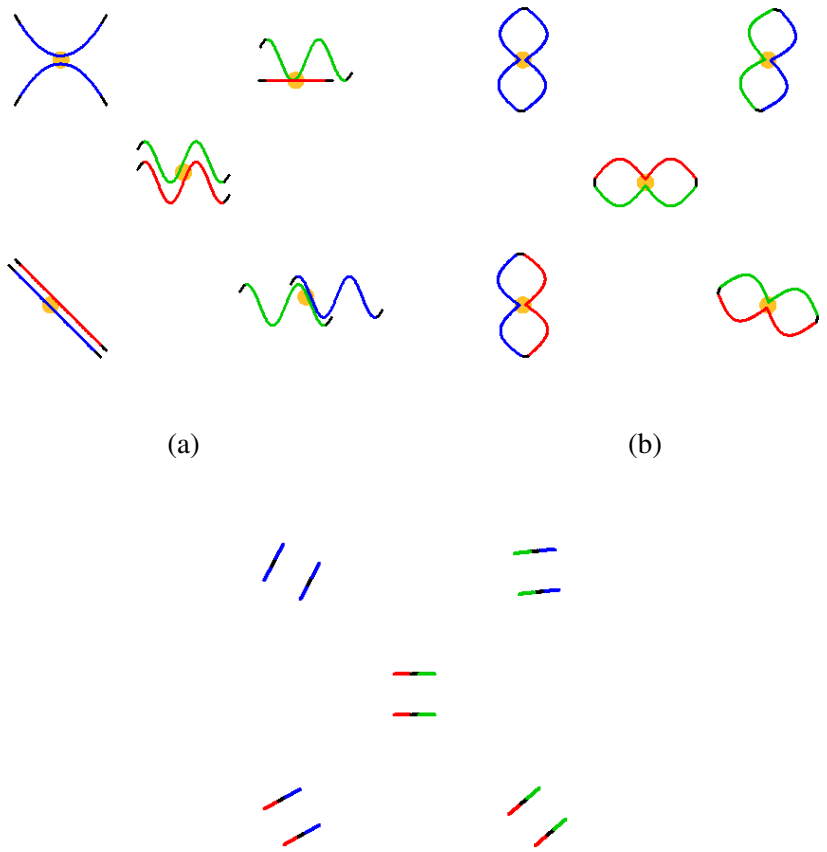
Figure 1.3: Illustration of the 4C process outlined in Section 1.2.2, beginning from Figure 1.2d. In Figure 1.3a a second round of shearing and ligation is applied to the ligated DNA segments, to produce DNA circles. In Figure 1.3b shows amplification of all the segments contacting the site of interest. Then the amplified DNA-DNA contacts are counted using large scaling sequencing. In the figures —, — and — denote DNA originating from different parts of the chromosome (or genome).



(a) (b)

(c)

Figure 1.4: Illustration of the 5C process outlined in Section 1.2.3, beginning from Figure 1.2d. In Figure 1.4a the new DNA segments are mixed with oligonucleotides, which partially overlap the different restriction sites of interest (green and red). In Figure 1.4b oligonucleotides corresponding to the new DNA segments are juxtaposed and ligated to produce new ligation products. In Figure 1.4c the new ligation products are simultaneously amplified and counted using large scale sequencing. In the figures —, — and — denote DNA originating from different parts of the chromosome (or genome). The oligonucleotides are denoted by - - - and - - -.



(a) (b)

(c)

Figure 1.5: Illustration of the Hi-C process outlined in Section 1.2.4, beginning from Figure 1.2b. In Figure 1.5a the restriction ends are filled using a biotin-labelled nucleotides. In Figure 1.5a blunt-end ligation is performed. In Figure 1.5c second round of purification and shearing, leaves the ligation junctions with the tagged biotin pull-down. The finally ligation junctions are mapped back to the genome using large scale sequencing. In the figures —, — and — denote DNA originating from different parts of the chromosome (or genome). The biotin pull-down is denoted by —.

# Chapter 2

# Multidimensional scaling

Multidimensional scaling (MDS) is the statistical field concerned with configuration recovery from distance (dissimilarity) matrices. MDS can be used for abstract interpretation such as mapping politicians on to an ideological axis (Diaconis et al., 2008), or can be used to recover a configuration of physical points such as the positions of cities within a country. MDS is a wide field with several different methods including metric MDS (Young and Householder, 1938; Torgerson, 1952); non-metric MDS (Shepard, 1962a,b; Kruskal, 1964a,b); maximum-likelihood MDS (Ramsay, 1977) and Bayesian MDS (Oh and Raftery, 2001; Hu et al., 2013). These MDS methods have been applied in a variety of disciplines such as psychology, ecology, political science and are beginning to play a role in inferring genome structure. This thesis uses metric and non-metric MDS on distance matrices estimated from Hi-C data to recover chromosome (or genome) configurations and seeks improvements to the process.

**Definition** A $(n \times n)$ matrix $\mathbf{D} = (d_{i,j})$ is a distance matrix if $d_{i,i} = 0 \forall\, i$, $d_{i,j} \geq 0 \,\forall\, i \neq j$, it is symmetric $d_{i,j} = d_{j,i} \,\forall i \neq j$, and satisfies the triangle inequality (2.1). **Definition** The triangle inequality states the distance between two points should be the shortest route

between those points

$$d_{i,j} \leq d_{i,a} + d_{a,j}. \tag{2.1}$$

Let $\mathbf{D} = (d_{i,j})$ be the $n \times n$ matrix of distances between $n$ points in $p$ dimensional space, where $\mathbf{X}$ is the $n \times p$ matrix of points. The $d_{i,j}$ can be any kind of distance (Cox and Cox (2000) page 11) or dissimilarly but here it is an Euclidean distance (2.2).

**Definition** Euclidean distance $d_{i,j}$ between points $\underline{x}_i = (x_{i,k})$ and $\underline{x}_j = (x_{j,k})$ is given by

$$d_{i,j} = \left( \sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2 \right)^{\frac{1}{2}} \tag{2.2}$$

**Definition** A distance matrix $\mathbf{D}$ is Euclidean, if it contains the Euclidean interpoint distances of the $n \times p$ configuration $\mathbf{X}$.

The Euclidean distance matrix $\mathbf{D}$ is a map of the configuration $\mathbf{X}$ and MDS can be used to recover this configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$, where $\hat{\mathbf{X}}$ is a $n \times p$ matrix of fitted points. If $\mathbf{D}$ contains no error then $\mathbf{X}$ and $\hat{\mathbf{X}}$ are identical up to some invariant transformation. If $\mathbf{D}$ contains error then it can be better described as an estimated or perturbed distance matrix $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$, where $\tilde{d}_{i,j} = d_{i,j} + \epsilon_{i,j}$ or $\tilde{d}_{i,j} = f(d_{i,j} + \epsilon_{i,j})$, where $\epsilon_{i,j}$ is some quantity of error $\epsilon \geq -d_{i,j}$ and $f(\ldots)$ is a monotonic function. The estimated or perturbed distance matrix $\tilde{\mathbf{D}}$ will likely of lost its Euclidean properties, but still contain a fuzzy map of the configuration $\mathbf{X}$. Multidimensional scaling will recover a configuration $\hat{\mathbf{X}}$, which is as close as possible the MDS method can get to the true configuration $\mathbf{X}$ up to some invariant transformation.

## 2.1 Metric multidimensional scaling

Metric MDS uses matrix algebra to recover a unique $n \times p$ fitted configuration $\hat{\mathbf{X}}$ from a distance matrix $\mathbf{D}$. The method of metric MDS begins with an element-wise transformation on $\mathbf{D}$ to obtain the $n \times n$ intermediate matrix $\mathbf{A} = (a_{i,j})$ where

$$a_{i,j} = -\frac{1}{2} d_{i,j}^2. \tag{2.3}$$

Then $\mathbf{A}$ is centred to obtain the centred inner-product matrix $\mathbf{B} = (b_{i,j})$,

$$\begin{aligned} \mathbf{B} &= \mathbf{HAH} \qquad (2.4)\\ b_{i,j} &= a_{i,j} - n^{-1} \sum_{j=1}^{n} a_{i,j} - n^{-1} \sum_{i=1}^{n} a_{i,j} + n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j}. \end{aligned}$$

where $\mathbf{H}$ is the $n \times n$ centring matrix

$$\mathbf{H} = \mathbf{I}_n - n^{-1} \underline{1}_n \underline{1}_n^T \tag{2.5}$$

The centred inner-product matrix now has the structure of

$$\mathbf{B} = (\mathbf{HX})(\mathbf{HX})^T. \tag{2.6}$$

The eigenvalue decomposition of $\mathbf{B}$ gives the eigenvalues $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p, 0, \ldots, 0)$ and corresponding eigenvectors $\mathbf{\Gamma} = (\gamma_{i,k})$, which are used to recover $\hat{\mathbf{X}}$:

$$\begin{aligned} \mathbf{B} &= \mathbf{\Gamma \Lambda \Gamma}^T \qquad (2.7)\\ \hat{\mathbf{X}} &= \mathbf{\Gamma \Lambda}^{\frac{1}{2}} \qquad (2.8)\\ x_{i,k} &= \gamma_{i,k} \lambda_k^{\frac{1}{2}}. \end{aligned}$$

The fitted configuration $\hat{\mathbf{X}}$ has the same configuration as $\mathbf{X}$ after centring, reflection and rotation:

$$\hat{\mathbf{X}} = (\mathbf{HX})\mathbf{R} \tag{2.9}$$

where $\mathbf{R}$ is a $p \times p$ orthogonal reflection and rotation matrix. Reconstructing $\mathbf{B}$ from $\hat{\mathbf{X}}$

$$\begin{aligned}\mathbf{B} &= \hat{\mathbf{X}}\hat{\mathbf{X}}^T \\ &= (\mathbf{HX})\mathbf{R}\mathbf{R}^T(\mathbf{HX})^T \\ &= (\mathbf{HX})(\mathbf{HX})^T \end{aligned} \tag{2.10}$$

because $\mathbf{R}\mathbf{R}^T = \mathbf{I}$ we obtain (2.6). Further information on metric MDS can be found in Cox and Cox (2000) pages 31-58 and Mardia et al. (1979) pages 397-402.

When fitting an estimated or perturbed distance matrix $\tilde{\mathbf{D}}$ into Euclidean space, the metric MDS produces a fitted configuration which is an approximation to the original, $\hat{\mathbf{X}} \approx (\mathbf{HX})\mathbf{R}$. This is because the perturbed centred inner product matrix $\tilde{\mathbf{B}}$ found from $\tilde{\mathbf{D}}$ using (2.4) differs from the true $\mathbf{B}$, and the subsequent fitted eigenvalues $\hat{\mathbf{\Lambda}} = (\hat{\lambda}_k)$ and fitted eigenvectors $\hat{\mathbf{\Gamma}} = (\hat{\underline{\gamma}}_k)$ from applying (2.7) to $\tilde{\mathbf{B}}$ differ from the original $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$. The number of non-zero eigenvalues in $\hat{\mathbf{\Lambda}}$ will differ from the original $\mathbf{\Lambda}$ as spurious positive non-zero eigenvalues are produced, $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_q > 0$ where $q \neq p$ (usually $q > p$), and if the Euclidean properties of $\tilde{\mathbf{D}}$ are broken some spurious negative fitted eigenvalues are produced $0 \geq \hat{\lambda}_{q+2} \geq \ldots \geq \hat{\lambda}_n$. The negative eigenvalues correspond to complex dimensions, recruited to absorb over fitted distances in the real dimensions. At low levels of perturbation in $\tilde{\mathbf{D}}$, $\hat{\lambda}_k \approx \lambda_k$ for $k = 1, \ldots, p$, at this point there is a fuzzy map of $\mathbf{X}$ held within $\tilde{\mathbf{D}}$. As the perturbation increases in $\tilde{\mathbf{D}}$ the fitted eigenvalues grow more dissimilar to the original eigenvalues until $\hat{\lambda}_k \neq \lambda_k$ for $k = 1, \ldots, p$, at this point the fuzzy map of $\mathbf{X}$ held within $\tilde{\mathbf{D}}$ has been completely erased.

Mardia (1978) uses $\lambda_k$ to calculate the percentage of information in $\mathbf{D}$ that is projected into the $k^{th}$ dimension:

$$\theta_k = \frac{|\lambda_k|}{\sum_{i=1}^{n} |\lambda_i|} \times 100\%, \tag{2.11}$$

then $\theta_k$ can be summed to give the percentage of information projected into the first $p$ dimensions

$$\theta_{1:p} = \sum_{k=1}^{p} \theta_k. \tag{2.12}$$

For example the distance matrix $\mathbf{D}$ for the configuration $\mathbf{X}$, where the number of points in $\mathbf{X}$ are even and are positioned equally spaced on the circumference of a circle. The percentages of information (2.11) from $\mathbf{D}$ projected into the first two dimensions are $\theta_1 = 50\%$, $\theta_2 = 50\%$ and $\theta_{1:2} = 100\%$, this is because the information in $\mathbf{D}$ can be equally projected into the first two dimensions. Now applying some small perturbation to $\mathbf{D}$ to give $\tilde{\mathbf{D}}$, the percentages of information from $\tilde{\mathbf{D}}$ projected into the first two dimensions are $\theta_1 = 47.8\%$, $\theta_2 = 43.7\%$ and $\theta_{1:2} = 91.5\%$. The total information has dropped from $\theta_{1:2} = 100\%$ to $\theta_{1:2} = 91.5\%$ after perturbation, this is because spurious non-zero eigenvalues are produced which projects information into additional dimensions.

In practice, metric MDS can be performed using the `cmdscale` function in the statistical software R, which requires a distance matrix `D` and the number of dimensions `k` (if unspecified `k=2`) and returns the metric MDS configuration. Additional input can be included in `cmdscale` with further instruction found on the internet.

## 2.2 Non-metric multidimensional scaling

Non-metric MDS uses optimization to recover a $n \times p$ fitted configuration $\hat{\mathbf{X}}$ representative of a dissimilarity matrix $\boldsymbol{\Delta} = (\delta_{i,j})$, where the Euclidean properties are relaxed and

$$\delta_{i,j} = f(d_{i,j}),$$

where $f$ is a monotonic function such that $d_{i,j} < d_{u,v}$ implies $\rightarrow$ $\delta_{i,j} < \delta_{u,v}$. Non-metric MDS orders the off-diagonal $\delta_{i,j}$ such that

$$\delta_{i_1,j_1} \leq \ldots \leq \delta_{i_m,j_m}$$

where $m = \frac{n(n-1)}{2}$, and seeks a fitted configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$ in $p$ dimensions, such that the fitted distances $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ (found by substituting $\hat{x}_{i,k}$ and $\hat{x}_{j,k}$ in for $x_{i,k}$ and $x_{j,k}$ in (2.2)) preserves the ordering

$$\hat{d}_{i_1,j_1} \leq \ldots \leq \hat{d}_{i_m,j_m}.$$

To measure how the ordering of the elements between $\boldsymbol{\Delta}$ and $\hat{\mathbf{D}}$ differs, the squared stress is used

$$S_p^2(\hat{\mathbf{X}}) = \frac{\sum_{i<j}(d^*_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} \hat{d}^2_{i,j}}, \tag{2.13}$$

where the $p$ denotes the number of dimensions $\hat{\mathbf{X}}$ is fitted into, the $d^*_{i,j}$ are the $\delta_{i,j}$ monotonically regressed (Izenman (2009) pages 493 - 497) onto $\hat{d}_{i,j}$, this allows the $d^*_{i,j}$ to share the same ordering as $\delta_{i,j}$ but are relative to $\hat{d}_{i,j}$ in size. The denominator of (2.13) makes $S_p^2(\hat{\mathbf{X}})$ invariant to uniform scaling. The square root of (2.13) is taken to give the

stress of fit statistic

$$S_p(\hat{\mathbf{X}}) = \left( \frac{\sum_{i<j} (d^*_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} \hat{d}^2_{i,j}} \right)^{\frac{1}{2}}. \tag{2.14}$$

Starting with an initial configuration $\hat{\mathbf{X}}^{(0)}$ in $p$ dimensional space, non-metric MDS iteratively adjusts $\hat{\mathbf{X}}$ and recalculates $S_p(\hat{\mathbf{X}})$, using the method of steepest descent to minimize $S_p(\hat{\mathbf{X}})$. When $S_p(\hat{\mathbf{X}}) = 0\%$, the fitted configuration $\hat{\mathbf{X}}$ is identical to the original configuration after some invarient transformation $\hat{\mathbf{X}} = b(\mathbf{HX})\mathbf{R}$, where $b > 0$ is a scaling constant, $\mathbf{H}$ is the centring matrix (2.5) and $\mathbf{R}$ is a $p \times p$ orthogonal reflection and rotation matrix.

When fitting a perturbed or estimated dissimilarity matrix, such that the ordering of perturbed dissimilarities $\tilde{\mathbf{\Delta}} = (\tilde{\delta}_{i,j})$ does not entirely reflect the true ordering in $d_{i,j}$ then $S_p(\hat{\mathbf{X}}) \neq 0\%$. In Mardia et al. (1979) page 414 provides the following guide to assess stress: $S_p(\hat{\mathbf{X}}) \geq 20\%$ is poor; $S_p(\hat{\mathbf{X}}) = 10\%$ is fair; $S_p(\hat{\mathbf{X}}) \leq 5\%$ is good and $S_p(\hat{\mathbf{X}}) = 0\%$ is perfect. Non-metric MDS does not provide a unique solution and can locate a local minimum to $S_p(\hat{\mathbf{X}})$ instead of a global minimum. More information on metric MDS can be found in Cox and Cox (2000) pages 61-90 and Mardia et al. (1979) pages 413-415.

In practice non-metric MDS is performed using `isoMDS` in R, which requires a dissimilarity matrix $\mathbf{\Delta}$ and the number of dimensions `k` (if unspecified `k=2`) and returns a configuration and corresponding stress. Additional input can be included in `isoMDS`. The initial configuration $\hat{\mathbf{X}}^{(0)}$ is the metric solution although any initial configuration can be specified. To avoid `isoMDS` locating a local minimum, the fitting process is repeated 100 times using $\hat{\mathbf{X}}^{(0)}$ with elements generated from the uniform distribution $\hat{x}^{(0)}_{i,k} \sim \mathrm{U}(-0.5, 0.5)$, and once using the fitted configuration $\hat{\mathbf{X}}^{(0)}$ from metric MDS, finally the $\hat{\mathbf{X}}$ with the smallest $S_p(\hat{\mathbf{X}})$ (2.14) is chosen as the final configuration.

### 2.2.1 Scaling by majorizing a complicated function

As an alternative MDS method the scaling by majorizing a complicated function (SMACOF) algorithm could be used (De Leeuw, 2011). The SMACOF algorithm uses an iterative procedure to produce a fitted configuration from a matrix estimated distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$, where the fitted configuration minimizes the stress function

$$\sigma(\hat{\mathbf{X}}) = \sum_{i<j} (\tilde{d}_{i,j} - \hat{d}_{i,j})^2. \tag{2.15}$$

The stress function (2.15) is majorized and a fitted configuration is found which minimizes the majorization. A function $p(x)$ is majorized when a simple support function $q(x, y)$ is used instead of $p(x)$. The properties the support function $q(x, y)$ being $p(x) \leq q(x, y)$ for all $x$ in the domain of $p$ and $p(y) = q(y, y)$ for all $y$ in the domain of $q$.

The SMACOF algorithm when applied to dissimilarity data gives similar results to when non-metric MDS is used. Therefore the SMACOF algorithm was not used for further analysis.

## 2.3 Procrustes shape distance

The Procrustes shape distance is the ordinary sum of squares between two configurations $\text{OSS}(\mathbf{X}, \hat{\mathbf{X}})$ after optimal reflection, rotation, scaling and translation. If $\mathbf{X}$ and $\hat{\mathbf{X}}$ are centred on the origin, translation is not required and the $\text{OSS}(\mathbf{X}, \hat{\mathbf{X}})$ can be written as

$$\text{OSS}(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{k=1}^{p} \sum_{i=1}^{n} \left( \mathbf{X} - \hat{b}\hat{\mathbf{R}}^T \hat{\mathbf{X}} \right)^2 \tag{2.16}$$

where $\underline{x}_i$ and $\hat{\underline{x}}_i$ are coordinate vectors for the $i^{\text{th}}$ point, $\hat{b} > 0$ is a scaling constant, and $\hat{\mathbf{R}}$ is an orthogonal $p \times p$ reflection and rotation matrix. To find $\hat{b}$ and $\hat{\mathbf{R}}$, let

$$\mathbf{Z} = \hat{\mathbf{X}}^T \mathbf{X}. \tag{2.17}$$

Applying the singular value decomposition (SVD) to $\mathbf{Z}$ gives

$$\mathbf{Z} = \mathbf{V}\mathbf{Q}\mathbf{U}^T \tag{2.18}$$

where $\mathbf{V}$ and $\mathbf{U}$ are $p \times p$ orthogonal matrices and $\mathbf{Q}$ is a diagonal matrix of singular values. Then the values for $\hat{b}$ and $\hat{\mathbf{R}}$ are

$$\hat{\mathbf{R}} = \mathbf{V}\mathbf{U}^T \text{ and } \hat{b} = \frac{\text{tr}(\mathbf{Q})}{\text{tr}(\hat{\mathbf{X}}\hat{\mathbf{X}}^T)}.$$

A deeper description of the Procrustes distances is given in Cox and Cox (2000) Chapter 5 and Mardia et al. (1979), pages 416-419.

## 2.4 Horseshoe effect

The horseshoe effect is the tendency for multidimensional scaling to fit a horseshoe-shaped configuration, from data which do not necessarily arise from such a configuration. Characteristics of horseshoe configurations are points arranged to form a horseshoe in the first and second dimensions, and the points taking a cubic polynomial arrangement in the first and third dimensions.

The horseshoe effect in MDS has been observed in the archaeological, ecological, political and psychological sciences. Several studies have proposed methods to correct for horseshoes, Hill (1974) proposed detrended correspondence analysis to unfold the horseshoe, and Podani and Miklos (2002) investigated raising $d_{i,j}$ to different powers in

(2.3) to find some detrending effect, but it also increased the magnitude of the negative eigenvalues.

Mardia et al. (1979) page 412 suggests the horseshoe effect is caused by accurate local distances and inaccurate medium to large distances, causing medium and large distances to merge and the metric MDS to bring distant points closer together. This decrease in accuracy with increasing distance was investigated by Diaconis et al. (2008). Diaconis et al. transformed Euclidean distances so local distances remained accurate and large distances lost accuracy, showing that the eigenvectors from the $\mathbf{B}$ followed trigonometric functions and produced horseshoes. De Leeuw (2008) supported these findings explaining the nature of the distance matrices that Diaconis et al. had used would inevitabily result in horseshoes. De Leeuw (2008) further points out that applying metric MDS to toeplitz matrices will produce horseshoes. The toeplitz matrix is where the elements of the super-diagonals and sub-diagonals of the matrix, are a constant value. Figure 2.1 gives an example of a $5 \times 5$ symmetric Toeplitz matrix. The horseshoe effect appears to be a

$$\begin{pmatrix} 0 & a & b & c & d \\ a & 0 & a & b & c \\ b & a & 0 & a & b \\ c & b & a & 0 & a \\ d & c & b & a & 0 \end{pmatrix}$$

Figure 2.1: Example of a $5 \times 5$ Toeplitz matrix. The matrix is symmetric. The elements on each sub-diagonal and corresponding super-diagonal of the matrix equal.

product of decreased accuracy in medium and large distances, this is most common when metric MDS provides an abstract interpretation of object positioning. Poor judgement in measuring dissimilarities between objects provides the conditions for horseshoes, in contrast to the estimation where distances between physical objects are measured accurately.

An example of the horseshoe effect can be given by artificially distorting a distance

matrix, using a distortion similar to the distortion used by Diaconis et al.. Take a configuration $\mathbf{X}$ of 20 equally spaced points on the unit line, with interpoint distances $\mathbf{D}$. The distances are distorted to give distorted distances $\mathbf{D}^* = (d^*_{i,j})$

$$d^*_{i,j} = 1 - \exp(-5 \times d_{i,j}) \tag{2.19}$$

The distortion (2.19) merges the medium distances of $\mathbf{D}$ into large distances (displayed in Figure 2.2), causing the confusion required for the horseshoe effect in the fitted configuration. The $\mathbf{D}^*$ are then fitted into three dimensional Euclidean space using metric MDS, to give the fitted configuration $\hat{\mathbf{X}}^*$. Figure 2.3 displays the fitted configuration $\hat{\mathbf{X}}^*$. In the first and second dimensions the configuration looks like a parabola with involution at the ends. In the first and third dimension the configuration looks to have a cubic polynomial arrangement with involution at the ends.

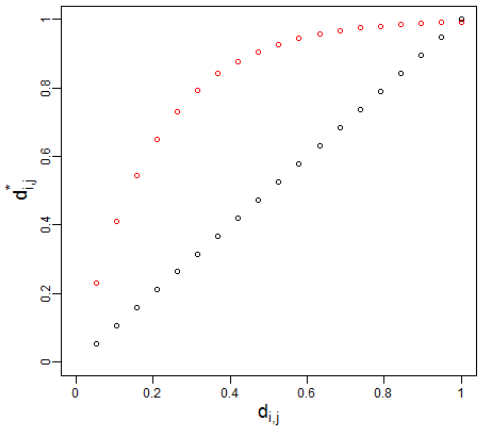



Figure 2.2: Distorted distances for the horseshoe example. The black circle ○ denotes the sizes of the distances $d_{i,j}$, between twenty equally spaced points on a unit line. The red circle ○ denotes the sizes of the distorted distances $d^*_{i,j}$, where (2.19) gives the distortion.

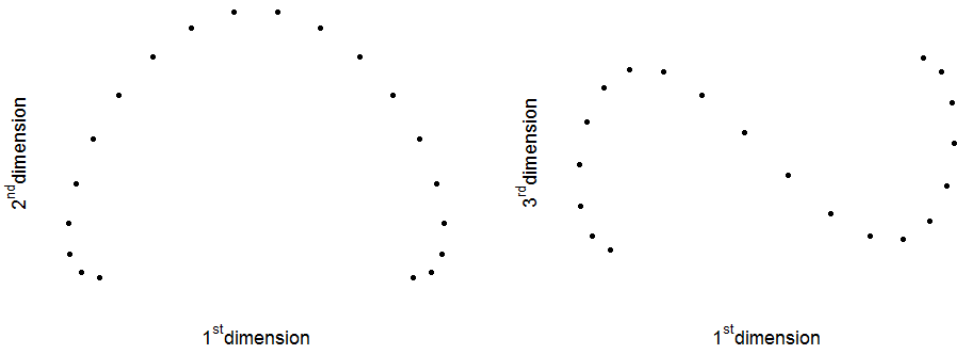



Figure 2.3: Perspectives of the fitted configuration for the horseshoe example. The fitted configuration $\hat{\mathbf{X}}^*$ is found by fitting the distorted distance matrix $\mathbf{D}^*$ (2.19) into three dimensional Euclidean space using metric multidimensional scaling.

# Chapter 3

# Investigation into metric multidimensional scaling

This chapter investigates some of the properties of metric multidimensional scaling (MDS). The first part manually performs metric MDS on $2\times2$ and $3\times3$ distance matrices, to glean information on how distances are arranged in the equations for the eigenvalues; eigenvectors and the fitted coordinates. The latter part distorts distances to the point where the distance matrix is no longer Euclidean and applies metric MDS, to observe how the fitted eigenvalues are disrupted by the distortion.

## 3.1 Fitting small distance matrices using metric MDS

Manually fitting $2 \times 2$ and $3 \times 3$ sized distance matrices, can give the fitted eigenvalues; eigenvectors and the points of the fitted configuration in terms of the distances, opening a small window on how distance affects the fitted configuration.

### 3.1.1 $2 \times 2$ distance matrix

The process of fitting a $2 \times 2$ distance matrix $\mathbf{D}$ with metric MDS is performed manually, where the two points are separated by a distance $d$. Let $\mathbf{X} = (\underline{x}_{(1)}, \underline{x}_{(2)})$ where $\underline{x}_{(k)} = (x_{1,k}, x_{2,k})^T$ are the point coordinates in dimension $k = 1, 2$. $\mathbf{D}$ first produces the centred inner product matrix $\mathbf{B}$ using (2.4),

$$\mathbf{D} = \begin{pmatrix} 0 & d \\ d & 0 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} \frac{d^2}{4} & \frac{-d^2}{4} \\ \frac{-d^2}{4} & \frac{d^2}{4} \end{pmatrix}.$$

The eigenvalue decomposition is applied to $\mathbf{B}$

$$|\mathbf{B} - \lambda \mathbf{I}_2| = \lambda^2 - 2\lambda \frac{d^2}{4} \tag{3.1}$$

Solving (3.1) for eigenvalues gives,

$$\lambda_1 = \frac{d^2}{2} \text{ and } \lambda_2 = 0,$$

and corresponding eigenvectors

$$\underline{\gamma}_1 = \left( \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}} \right)^T \text{ and } \underline{\gamma}_2 = \left( \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T.$$

$\mathbf{X}$ is reconstructed using $\underline{x}_{(k)} = \underline{\gamma}_k \lambda_k^{\frac{1}{2}}$ to give

$$\underline{x}_{(1)} = \left( \frac{d}{2}, \frac{-d}{2} \right)^T \text{ and } \underline{x}_{(2)} = (0, 0)^T \tag{3.2}$$

The solution (3.2) is trivial as the points only require placing a distance $d$ apart to be recovered, although it does serve as a warm-up for the $3 \times 3$ distance matrix.

### 3.1.2 $3 \times 3$ distance matrix

The $3 \times 3$ case is similar to the $2 \times 2$, case let $\mathbf{X} = (\underline{x}_{(1)}, \underline{x}_{(2)}, \underline{x}_{(3)})$, where $\underline{x}_{(k)}$ gives the coordinates of points in dimension $k$. The $3 \times 3$ distance matrix $\mathbf{D}$ first produces the centred inner product matrix $\mathbf{B}$ using (2.4),

$$\mathbf{D} = \begin{pmatrix} 0 & a & b \\ a & 0 & c \\ b & c & 0 \end{pmatrix} \text{ and}$$

$$\mathbf{B} = \frac{1}{18} \begin{pmatrix} 4a^2 + 4b^2 - 2c^2 & -5a^2 + b^2 + c^2 & a^2 - 5b^2 + c^2 \\ -5a^2 + b^2 + c^2 & 4a^2 - 2b^2 + 4c^2 & a^2 + b^2 - 5c^2 \\ a^2 - 5b^2 + c^2 & a^2 + b^2 - 5c^2 & -2a^2 + 4b^2 + 4c^2 \end{pmatrix}.$$

Applying the eigenvalue decomposition to $\mathbf{B}$ gives

$$\begin{aligned} |\mathbf{B} - \lambda \mathbf{I}_3| = &\left( \frac{-1}{6}(a^2b^2 + a^2c^2 + b^2c^2) + \frac{1}{12}(a^4 + b^4 + c^4) \right) \lambda \\ &+ \frac{1}{3}(a^2 + b^2 + c^2)\lambda^2 - \lambda^3, \end{aligned} \tag{3.3}$$

solving (3.3) for the eigenvalues gives

$$\begin{aligned} \lambda_1 &= \frac{1}{6}(a^2 + b^2 + c^2 + 2\sqrt{a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2}); \\ \lambda_2 &= \frac{1}{6}(a^2 + b^2 + c^2 - 2\sqrt{a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2}) \\ \text{and } \lambda_3 &= 0. \end{aligned} \tag{3.4}$$

To find the corresponding eigenvectors, $\mathbf{B}$ is rotated using a Helmert rotation matrix $\mathbf{R}$

$$\mathbf{R} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{6}}. \end{pmatrix}$$

$$\mathbf{R}^T\mathbf{B}\mathbf{R} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{b^2}{2} & \frac{(-a^2+c^2)}{2\sqrt{3}} \\ 0 & \frac{(-a^2+c^2)}{2\sqrt{3}} & \frac{(2a^2-b^2+2c^2)}{6} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \sigma^2_{1,1} & \sigma^2_{1,2} \\ 0 & \sigma^2_{1,2} & \sigma^2_{2,2} \end{pmatrix} \tag{3.5}$$

giving a $2 \times 2$ symmetric matrix nested within a $3 \times 3$ null matrix. (3.5) can be solved using a formula provided by Mardia et al. (1979), page 246, exercise 8.1.1 to give the eigenvectors of $\mathbf{R}^T\mathbf{B}\mathbf{R}$,

$$\underline{\phi}_1 = \begin{pmatrix} 0 \\ \sigma^2_{2,2} - \sigma^2_{1,1} + \Theta \\ -2\sigma^2_{1,2} \end{pmatrix} \text{ and } \underline{\phi}_2 = \begin{pmatrix} 0 \\ 2\sigma^2_{1,2} \\ \sigma^2_{2,2} - \sigma^2_{1,1} + \Theta \end{pmatrix} \tag{3.6}$$

where $\Theta = \sqrt{(\sigma^2_{1,1} - \sigma^2_{2,2})^2 + 4\sigma^4_{1,2}}$. The rotation is reversed by pre-multiplying the eigenvectors (3.6) by $\mathbf{R}$ to recover the eigenvectors of $\mathbf{B}$

$$\underline{\gamma}_1 = \begin{pmatrix} b^2 - c^2 + \Delta \\ -a^2 + c^2 \\ a^2 - b^2 - \Delta \end{pmatrix} \quad \underline{\gamma}_2 = \begin{pmatrix} 2a^2 - b^2 - c^2 - \Delta \\ -a^2 + 2b^2 - c^2 + 2\Delta \\ -a^2 - b^2 + 2c^2 - \Delta \end{pmatrix} \text{ and } \underline{\gamma}_3 = \begin{pmatrix} \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{pmatrix} \tag{3.7}$$

where $\Delta = \sqrt{a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2}$. Finally $\mathbf{X}$ is reconstructed using $\underline{x}_{(k)} = \underline{\gamma}_k \lambda_k^{\frac{1}{2}}$ to give

$$\underline{x}_{(1)} = w_1 \begin{pmatrix} b^2 - c^2 + \Delta \\ -a^2 + c^2 \\ a^2 - b^2 - \Delta \end{pmatrix}$$

where

$$w_1 = \left( \frac{a^2 + b^2 + c^2 + 2(a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2)}{6((b^2 - c^2 + \Delta)^2 + (-a^2 + c^2)^2 + (a^2 - b^2 - \Delta)^2)} \right)^{\frac{1}{2}},$$

$$\underline{x}_{(2)} = w_2 \begin{pmatrix} 2a^2 - b^2 - c^2 - \Delta \\ -a^2 + 2b^2 - c^2 + 2\Delta \\ -a^2 - b^2 + 2c^2 - \Delta \end{pmatrix}$$

where

$$w_2 = \left( \frac{a^2 + b^2 + c^2 - 2(a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2)}{6((2a^2 - b^2 - c^2 - \Delta)^2 + (2b^2 - a^2 - c^2 + 2\Delta)^2 + (2c^2 - a^2 - b^2 - \Delta)^2)} \right)^{\frac{1}{2}}$$

and $\underline{x}_{(3)} = (0, 0, 0)^T$ since $\lambda_3 = 0$.

The constants $w_1$ and $w_2$ are a product of $\lambda_k^{\frac{1}{2}}$ and the eigenvector normalisation constant. The eigenvalues (3.4) and eigenvectors (3.7) explicity show how distances produce coordinates.

Equation (3.4) can be used to determine if the desired Euclidean properties of $\mathbf{D}$ are violated. Rearranging the equation for the second eigenvalues (3.4) gives

$$a^2 + b^2 + c^2 \geq 2\sqrt{a^4 + b^4 + c^4 - a^2b^2 - a^2c^2 - b^2c^2},$$

and if this inequality holds then $\mathbf{D}$ is Euclidean.

## Illustration

Consider three points spaced one unit apart on a line for which

$$\mathbf{D} = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \text{ and } \mathbf{D}^2 = \begin{pmatrix} 0 & 1 & 4 \\ 1 & 0 & 1 \\ 4 & 1 & 0 \end{pmatrix}.$$

Increase the distance between points $1$ and $2$ by adding some quantity $\epsilon$ to the squared distance gives

$$\mathbf{D}^2 = \begin{pmatrix} 0 & 1+\epsilon & 4 \\ 1+\epsilon & 0 & 1 \\ 4 & 1 & 0 \end{pmatrix}, \tag{3.8}$$

where $\epsilon \geq -1$ to avoid the squared distance becoming negative. The resulting eigenvalues using (3.4) are

$$\lambda_1 = \frac{1}{6}(6 + \epsilon + 2\Delta) \text{ and } \lambda_2 = \frac{1}{6}(6 + \epsilon - 2\Delta)$$

where $\Delta = \sqrt{(9 - 3\epsilon + \epsilon^2)}$. The $\lambda_2$ is of most interest and can be categorized into two states. The first state is when $0 \leq \epsilon \leq 8$, for which the Euclidean properties of $\mathbf{D}$ are intact because $\lambda_2 \geq 0$. The second state is when $-1 \leq \epsilon \leq 0$ or $\epsilon > 8$ for which the Euclidean properties are violated because $\lambda_2 < 0$. The second state is of most interest, here the second dimension is projected into complex space. The distance contribution of the second dimension acts to reduce the distance expansion in the first dimension, such that

$$d^*_{1,2} = \sqrt{(x_{1,1} - x_{2,1})^2 + i^2(x_{1,2} - x_{2,2})^2} = \sqrt{(x_{1,1} - x_{2,1})^2 - (x_{1,2} - x_{2,2})^2}.$$

## 3.2 Distortion

The eigenvalues from fitting the $3 \times 3$ distance matrix (3.4) quantify the robustness of $\mathbf{D}$ to distortion. Manually fitting a $4 \times 4$ distance matrix, would explicitly show how distances go to produce the eigenvalues and eigenvectors, unfortunately computation increases for each additional point added, making fitting the $4 \times 4$ distance matrix much more difficult than fitting the $3 \times 3$ distance matrix $\mathbf{D}$.

To investigate how robust larger $\mathbf{D} = (d_{i,j})$ are and gain insight into negative eigenvalues, distance matrices were deliberately distorted by adding $\epsilon$ to one of the $d_{i,j}$ (and the symmetric $d_{j,i}$) to give a distorted distance matrix $\mathbf{D}(\epsilon)$. These $\mathbf{D}(\epsilon)$ were then fitted into $k$ dimensional Euclidean space using metric MDS and the fitted eigenvalues were investigated.

### 3.2.1 Four point problem

The distance matrices $\mathbf{D}$ for three different four point configurations were investigated. One configuration being four points equally spaced on a straight lines (1D); another being four points on the corners of a square (2D) and another being four points on the corners of a tetrahedron (3D). The configurations were standardised so the shortest distance was a unit length. Then some measured error $\epsilon$ was added to one of the unit distances in each $\mathbf{D}$ where $-1 \leq \epsilon \leq 5$ to give the distorted distance matrix $\mathbf{D}(\epsilon)$. The matrix $\mathbf{D}(\epsilon)$ was fitted into $k$ dimensional Euclidean space (where $k$ are the dimensions corresponding to the positive eigenvalues), using metric MDS and the fitted eigenvalues and an assortment of distances were recorded as $\epsilon$ increased. The assortments of distances were the distorted distance $d^*_{i,j} = d_{i,j} + \epsilon$; fitted distorted distance $\hat{d}^*_{i,j}$; the sum of the undistorted distances $\sum_{(l,m)\neq(i,j)} d_{l,m}$ and the sum of the fitted undistorted distances $\sum_{(l,m)\neq(i,j)} \hat{d}_{l,m}$. The assortment provided information on the distances directly affected by distortion, and the

distances indirectly affected by distortion. Investigating the assortment of distances as $\epsilon$ increases, allows the observation of how the undistorted distances are coerced by the distorted distance. The eigenvalues in Figure 3.4 can be split into two states. In the first state all the eigenvalues are non-negative hence $\mathbf{D}(\epsilon)$ is still Euclidean. The first state exist briefly for the straight line ($\epsilon = 0$) and is more persistent for the square and tetrahedron. The persistence shows the distance matrices are robust to distortion up to a certain point. The undistorted fitted distances in the first state (Figure 3.5) remain equal to the undistorted distances hence $\mathbf{D}(\epsilon) \approx \hat{\mathbf{D}}(\epsilon)$, as metric MDS accommodates distortion in the available dimensions.

The second state is characterized by the negative eigenvalues as $\mathbf{D}(\epsilon)$ is no longer Euclidean. The second state can be subdivided into two substates when $-1 < \epsilon < 0$ and $\epsilon > 0$. In the first substate distortion reduces $d^*_{i,j}$ to the point where metric MDS has to increase it to fit with the undistorted distances, this can be observed in the plots of distances for the straight line and square.

The second substate is characterized by the principle positive and negative eigenvalues blowing up as $\epsilon$ becomes very large. The negative eigenvalues grow counter to the growth in the positive eigenvalues, they act to absorb the increase of undistorted and distorted fitted distances by the metric MDS. The large distorted distance coerces the metric MDS to increases the undistorted distances in fitting, making the undistorted fitted distances increase with distortion.

Distortion on $4 \times 4$ distance matrices display metric MDS accommodating distortion without breaking the Euclidean properties of $\mathbf{D}(\epsilon)$, up to a critical value of $\epsilon$, then for larger $\epsilon$ the distortion overwhelms the Euclidean properties of $\mathbf{D}(\epsilon)$ to create negative and positive eigenvalues. Eventually distortion becomes great enough that the lead dimension is determined by the distorted dimension and the contribution of the undistorted distances to the lead dimension are negligible.

$$\begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 1+\epsilon & 2 \\ 2 & 1+\epsilon & 0 & 1 \\ 3 & 2 & 1 & 0 \end{pmatrix}$$

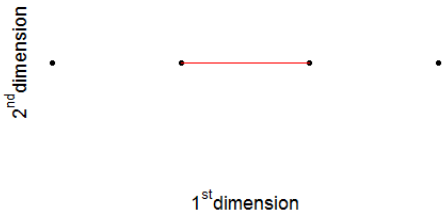



Figure 3.1: The distorted distance matrix $\mathbf{D}(\epsilon)$ and a illustration of the four point straight line. The red line — connects the two points in which the distance is distorted, by the addition of some quantity $\epsilon$ to the distance.

$$\begin{pmatrix} 0 & 1+\epsilon & 1 & \sqrt{2} \\ 1+\epsilon & 0 & \sqrt{2} & 1 \\ 1 & \sqrt{2} & 0 & 1 \\ \sqrt{2} & 1 & 1 & 0 \end{pmatrix}$$

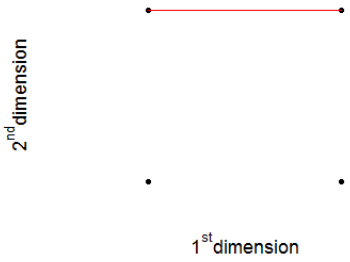



Figure 3.2: The distorted distance matrix $\mathbf{D}(\epsilon)$ and a illustration of the four points on the corners of a square. The red line — connects the two points in which the distance is distorted, by the addition of some quantity $\epsilon$ to the distance.

$$\begin{pmatrix} 0 & 1+\epsilon & 1 & 1 \\ 1+\epsilon & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

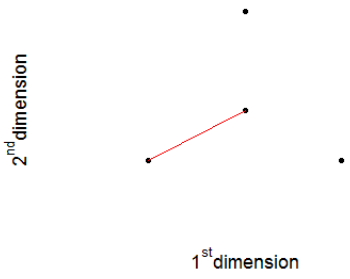
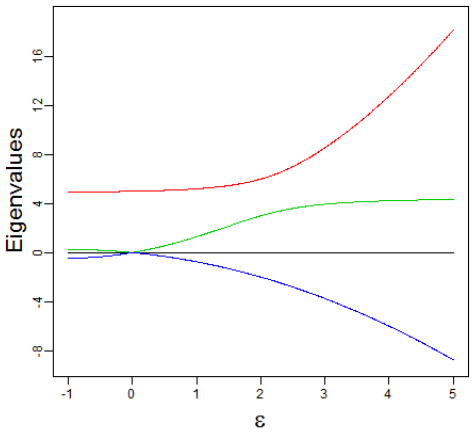
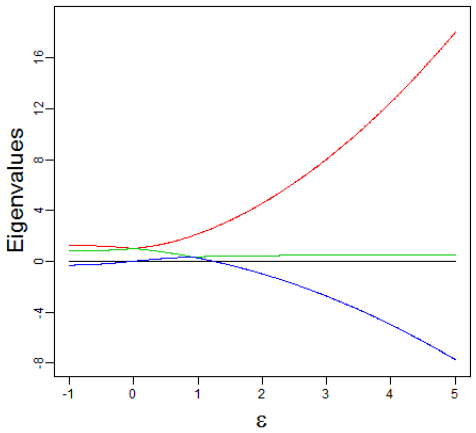


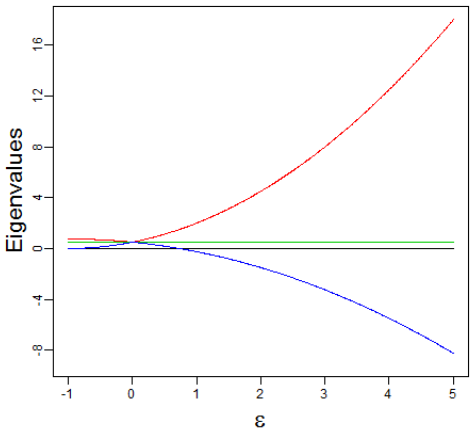

Figure 3.3: The distorted distance matrix $\mathbf{D}(\epsilon)$ and a illustration of the four points on the corners of a tetrahedron. The red line — connects the two points in which the distance is distorted, by the addition of some quantity $\epsilon$ to the distance.




(a) Straight line.

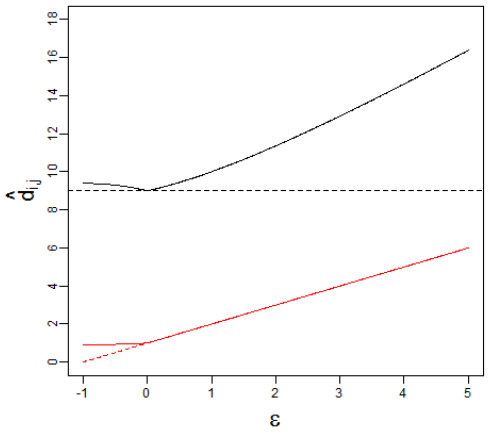



(b) Square.

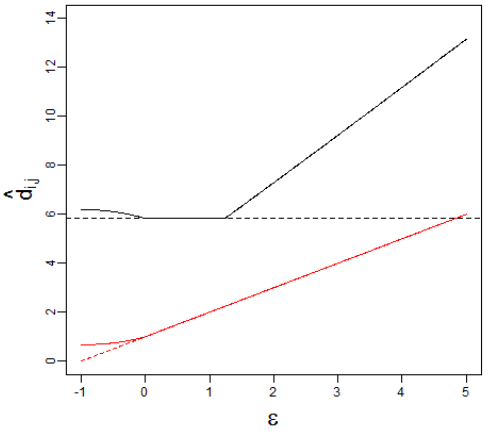



(c) Tetrahedron.

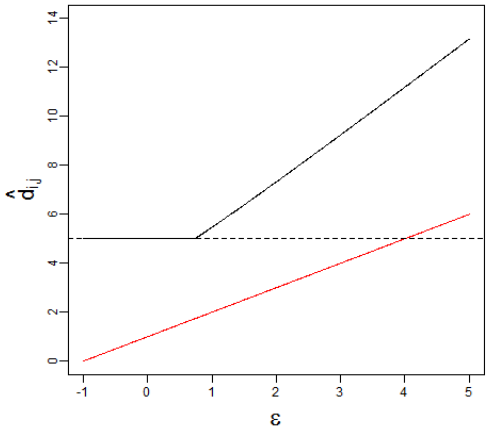
Figure 3.4: The fitted eigenvalues (2.7) from fitting the distorted distance matrices $\mathbf{D}(\epsilon)$ with metric MDS, as distortion increase. Top: four points on a straight line's fitted eigenvalues. Bottom left: four points on the corners of a square's fitted eigenvalues. Bottom right: four points on the corners of a tetrahedron's fitted eigenvalues. The red line — denotes the principal eigenvalue; the green line — the second eigenvalues; the blue line —, and the black line — the eigenvalue of fixed size zero.




(a) Straight line.




(b) Square.




(c) Tetrahedron.

Figure 3.5: The assortment of distances from fitting the distorted distance matrices $\mathbf{D}(\epsilon)$ with metric MDS, as distortion increase. Top: four points on a straight line's assortment of distances. Bottom left: four points on the corners of a square's assortment of distances. Bottom right: four points on the corners of a tetrahedron's assortment of distances. The dashed black line - - - gives the sum of the distances (before distortion) in $\mathbf{D}(\epsilon)$ excluding the distorted distances $\sum_{(l,m)\neq(i,j)} d_{l,m}$. The solid black line ___ gives the sum of the fitted distances $\hat{\mathbf{D}}(\epsilon)$ from (2.2) (where $p$ is the number of positive eigenvalues), excluding the fitted distorted distance $\sum_{(l,m)\neq(i,j)} \hat{d}_{l,m}$. The dashed red line - - - gives the distorted distance $d^*_{i,j}$; the solid red line ___ gives the fitted distorted distance $\hat{d}^*_{i,j}$ (2.2) (where $p$ is the number of positive eigenvalues). The black lines give information on the distances which are not directly distorted. The red lines give information on the distance which is directly distorted.

### 3.2.2 Lattices

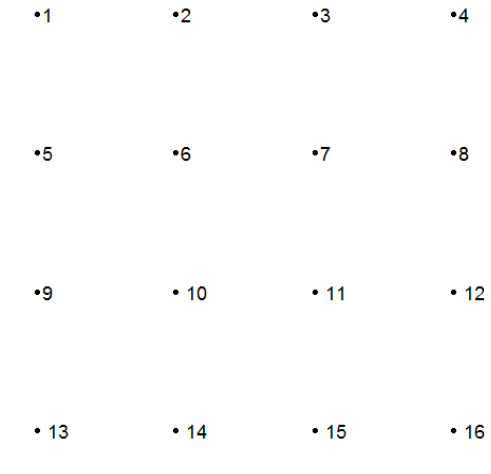
Distortion was applied to lattices of points, to investigate the spurious eigenvalues generated when the Euclidean properties of the distance matrix are broken.

Using a $4 \times 4$ lattice of points (Figure 3.6a), where points which are horizontal or vertical to each other are separated by a unit distance. The distance matrix for the lattice is distorted by adding $\epsilon$ to one or more of the distances. The fitted eigenvalues from fitting the distorted distance matrix with metric MDS are investigated. Three examples of distortion are given in Figure 3.6, to try convey the observations made through distortion. The first example distorts the distance between points 1 & 2. The second example distorts the distances between both 1 & 2 and 7 & 11. The final example distorts the distances between points 1 & 2; 7 & 11 and 15 & 16.

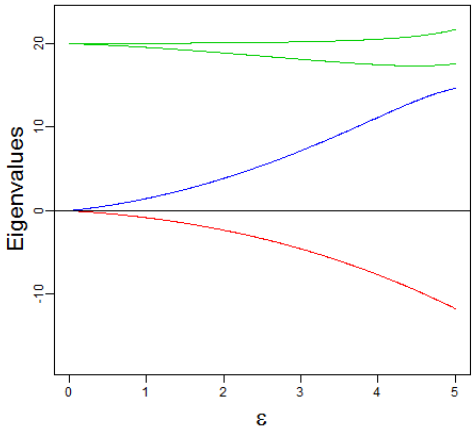
Through distorting different combinations of distances on different lattices, we found distorting the distance between any pair of given points $i_1$ & $j_1$, produces an additional positive and negative eigenvalue. Simultaneously distorting a further distance between points $i_2$ & $j_2$ where $(i_2, j_2) \neq (i_1, j_1)$, a further additional positive and negative eigenvalue is produced.

We found that distorting a small set of $p$ distances, where no point had more than one of its distances distorted, then $p$ additional positive and $p$ additional negative eigenvalues were produced. As $p$ increases this observation becomes weaker, such that fewer negative eigenvalues are produced. If $p = \frac{n}{2}$ so that each point has one distance distorted, there only $n - 3$ additional non-zero eigenvalues can be produced, so $\frac{n}{2}$ additional positive and $\frac{n}{2}$ additional negative eigenvalues will not be produced.

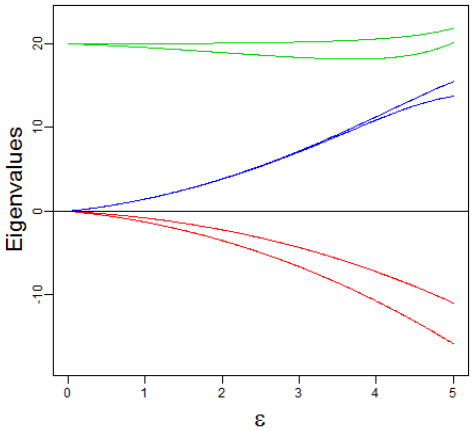
When distorting the distance between points $i_1$ & $j_1$ and $i_1$ & $j_2$ where $j_1 \neq j_2$, we observe an additional positive and negative eigenvalue, as removing point $i_1$ from the lattice and its distances from the distance matrix, gives a new lattice with Euclidean interpoint distances.



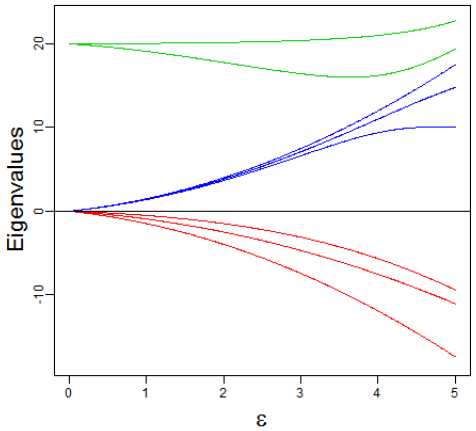

(a) Illustration of the lattice.




(b) One distance distorted.




(c) Two distances distorted.




(d) Three distances distorted.

Figure 3.6: Illustration of the $4 \times 4$ lattice, and the fitted eigenvalues (2.7) from distorting combinations of distances between the points on the lattice. Top left: illustration of the lattice with points numbered. Top right: fitted eigenvalues from distorting the distance between points 1 & 2. Bottom left: fitted eigenvalues from distorting the distance between points 1 & 2 and 7 & 11. Bottom right: fitted eigenvalues from distorting the distance between points 1 & 2, 7 & 11 and 15 & 16. The distances are distorted by simultaneously adding the same quantity $\epsilon$ to them. The green line — denotes the genuine positive fitted eigenvalues; the blue line — denotes the spurious positive fitted eigenvalues and the red line — denotes the spurious negative fitted eigenvalues.

### 3.2.3 Interpretation

Matrix algebra can provide some insight into the additional eigenvalues after distortion. Let $\mathbf{A} = (a_{i,j})$ be the intermediate matrix generated by (2.3) on the undistorted distance matrix $\mathbf{D}$. Let $\mathbf{E}$ be a sparse symmetric matrix with non-zero $\epsilon$ on the entries corresponding to where distortion is applied, where a point can only have one of its distances distorted, so each column or row has a maximum of one non-zero entry. The rank of $\mathbf{E}$ is then the number of distorted distances $e$. Summing $\mathbf{A}$ and $\mathbf{E}$ gives $\mathbf{A}(\epsilon) = \mathbf{A} + \mathbf{E}$ as the distorted intermediate matrix. Distortion applied to the intermediate matrix is easier to handle algebraically than distortion applied to the distance matrix. Centring $\mathbf{A}(\epsilon)$ using (2.4) gives $\mathbf{B}(\epsilon)$, the maximum rank of $\mathbf{B}(\epsilon)$, $\rho(\mathbf{B}(\epsilon))$ can be found using matrix algebra (Gentle, 2007).

$$\begin{aligned}
\rho(\mathbf{B}(\epsilon)) &= \rho(\mathbf{H}\mathbf{A}(\epsilon)\mathbf{H}) \\
&\leq \rho(\mathbf{H}\mathbf{A}\mathbf{H}) + \rho(\mathbf{H}\mathbf{E}\mathbf{H}) \\
&\leq \min(n-1, k) + \min(n-1, 2e) \\
&\leq k + 2e, \qquad (3.9)
\end{aligned}$$

where $k$ is the number of dimensions of the original configuration. The result (3.9) shows if one distance is distorted then there is a maximum of two additional non-zero eigenvalues.

How the centring of intermediate matrix (2.4) distributes $\epsilon$ about $\mathbf{B}(\epsilon)$ from a single distortion applied in $\mathbf{A}(\epsilon)$ can also be investigated, to provide a stepping point for further analysis of the distortions affect on the eigenvalues.

Let $a$ and $b$ be the points where some distortion added between them, and let $i$ and $j$ be points with no distortion. The distribution of the distortion through $\mathbf{B}(\epsilon)$ can be observed in Figure 3.7 when distortion is added between points 1 & 2. The level of distortion in

zones A; B; C and D in Figure 3.7 is given below.

$$\text{A: } b_{a,b} + \epsilon\left(1 - \frac{2}{n} + \frac{2}{n^2}\right), \tag{3.10}$$

$$\text{B: } b_{a,a} + 2\epsilon\left(\frac{1}{n^2} - \frac{1}{n}\right), \tag{3.11}$$

$$\text{C: } b_{a,j} + \frac{\epsilon}{n}\left(\frac{2}{n} - 1\right), \tag{3.12}$$

$$\text{D: } b_{i,j} + 2\frac{\epsilon}{n^2}. \tag{3.13}$$

The zones A; B and C have the strongest distortion as they are associated with the distorted points, with zones D having the weakest distortion.

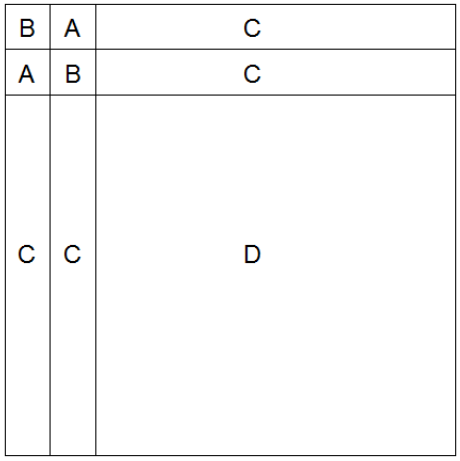



Figure 3.7: Illustration of how a single distortion in the intermediate matrix $\mathbf{A}$ (2.3) is distributed about the distorted centred inner product matrix $\mathbf{B}(\epsilon)$ (2.4). In this illustrations some quantity $\epsilon$ has been added to $\mathbf{A}$ elements $a_{1,2}$ and $a_{2,1}$ to give a distorted intermediate matrix $\mathbf{A}(\epsilon)$. The distorted intermediate matrix $\mathbf{A}(\epsilon)$ is centred using (2.3) to give $\mathbf{B}(\epsilon)$. The zones labelled on $\mathbf{B}(\epsilon)$ correspond to the distribution of distortion, zone A is given by (3.10); zone B is given by (3.11); zone C is given by (3.12) and zone D is given by (3.13).

Using the property $\text{trace}(\mathbf{B}) = \sum_{k=1}^{n} \lambda_k$, the total size change on $\mathbf{B}(\epsilon)$ eigenvalues $\lambda_k(\epsilon)$

for $k = 1, \ldots, n$ is

$$\sum_{k=1}^{n} \lambda_k(\epsilon) = \sum_{k=1}^{n} \lambda_k - 2\frac{\epsilon}{n}, \tag{3.14}$$

where $\lambda_k$ are the eigenvalues from the undistorted $\mathbf{B}$, and $\epsilon$ has been added to elements of $\mathbf{A}$.

## 3.3 Conclusion

Manually applying metric MDS to small distance matrices provides an insight into how the metric MDS works and shows how the computational difficulty of metric MDS increases as the number of point increases. Distortion in the four point problem offers insight into how the eigenvalues and distances adjust to accommodate distortion. Further interpretation of the distortion shows a maximum number of new eigenvalues produced by distortion, how $\epsilon$ is distributed in $\mathbf{B}(\epsilon)$ and how distortion effects the total size of the eigenvalues.

Manually applying metric MDS to small distance matrices and distortion provide a glimpse into the workings of metric MDS, but provide little help in elucidating chromosome or genome structure. Instead broader understanding of noise in the estimated distance matrix is required.

# Chapter 4

# Exploratory analysis

The first part of this chapter details our method of recovering an estimated chromosome configuration from the Hi-C (Lieberman-Aiden et al., 2009) contact frequency (count) matrix $\mathbf{M}$ of the Karyotypically normal human lymphoblastoid cell line (GM06990) at one megabase resolution. The first part of the chapter also details our interpretation and pre-processing of the data to improve the configuration. The latter part of the chapter details our pre-processing of the global count matrix, which contains the interchromosomal and intrachromosomal counts for the twenty two chromosome pairs and the XX chromosome pair in the GM06990 cell line, then applying the methods of estimated chromosome configuration recovery to estimating the genome configuration.

Let $\mathbf{X} = (x_{i,k})$ be the $n \times 3$ average chromosome configuration where $x_{i,k}$ for $k = 1, 2, 3$ is the centre of the megabase interval $i$. Interpoint distances of $\mathbf{X}$ are recorded in a $n \times n$ Euclidean distance matrix $\mathbf{D} = (d_{i,j})$ (2.2). The Euclidean distances have an unknown relationship with the hypothetical $n \times n$ expected chromosome count matrix $\mathbf{U} = (\mu_{i,j})$. The matrix $\mathbf{U}$ contains the hypothetical expected intrachromosomal counts between two megabase intervals or the same megabase interval. The observed chromosome count matrix $\mathbf{M} = (m_{i,j})$ contains random counts which can be modelled by some count distribution with mean $E(m_{i,j}) = \mu_{i,j}$ and variance var$(m_{i,j}) = \rho\mu_{i,j}$, where $\rho$ denotes

the level of dispersion in the counts $\rho \geq 1$. The matrix $\mathbf{M}$ can be regarded as a matrix of proximities, detailing how close and entwined megabase intervals are, the larger $m_{i,j}$ the greater the probability that the two megabase intervals $i$ and $j$ are spatially close. The matrix $\mathbf{M}$ can be transformed into an estimated distance matrix $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$ using a count to distance transform function $f(\ )$, where $f(\ )$ tries to emulate the relationship between $\mathbf{D}$ and $\mathbf{U}$ such that $f(\mu_{i,j}) \approx d_{i,j}$. Metric and non-metric multidimensional scaling (MDS) can be used to fit $\tilde{\mathbf{D}}$ into three-dimensional Euclidean space to give an estimated chromosome configuration $\hat{\mathbf{X}} = (\hat{x}_{i,j})$. The configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$ can be used to give fitted distances $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ by inputting the $\hat{x}_{i,k}$ into (2.2) with $p = 3$, and the fitted counts $\hat{\mathbf{U}} = (\hat{\mu}_{i,j})$ by inverting the count to distance transform $\hat{\mu}_{i,j} = f^{-1}(\hat{d}_{i,j})$. The average chromosome configuration $\mathbf{X}$; the true interpoint Euclidean distances $\mathbf{D}$; expected intrachromosomal counts $\mathbf{U}$ and the relationship between counts and distances are unknown; and the objective is to estimate $\mathbf{X}$ from $\mathbf{M}$.

## 4.1 Transform functions and measures of fit

This section outlines the count to distance transform function $f(\ )$ and the measures of fit used to assess $\hat{\mathbf{X}}$. These are the tools used with MDS to recover $\hat{\mathbf{X}}$.

### 4.1.1 Transform functions

The $m_{i,j}$ are noisy measures of proximity which require transforming into $\tilde{d}_{i,j}$, before MDS can be used to recover $\hat{\mathbf{X}}$, to do this a count to distance transform function $f(\ )$ is used. To produce $f(\ )$ the proximity properties of $\mathbf{M}$ are used as a guide.

The first property is the relationship between counts and distances should be

monotonically decreasing,

$$f(a) \leq f(b) \, \forall \, a \geq b. \tag{4.1}$$

The monotonically decreasing property comes from the intuition that megabase intervals of the chromosome spatially distant are likely to produce low counts, whereas spatially close megabase intervals are likely to produce large counts. This relationship is observed between count size and genomic distance: megabase intervals $i$ and $j$ which are genomically close have large $m_{i,j}$ whereas megabase intervals genomically distant have small $m_{i,j}$. The genomic distance measured from the centre of the megabase interval, is the measure of how many base pairs separate megabase intervals. The megabase interval $i$ and megabase interval $j$ have a genomic distance of $|i - j|$ megabases (Mb), whereas megabase interval $i$ and megabase interval $i+1$ abut each other and have genomic distance of 1 Mb.

The second property is that distance should tend to zero, as count size tends to infinity

$$f(\mu_{i,j}) \rightarrow 0 \text{ as } \mu_{i,j} \rightarrow \infty. \tag{4.2}$$

The second property can be observed in the counts a megabase interval makes with itself, found on the diagonal of $\mathbf{M}$. Although $m_{i,i} \neq \infty$ they do represent the count size required to give a distance of zero $\tilde{d}_{i,j} = 0$.

The third property is the rate of decrease in $f(\mu_{i,j})$ should tend to zero, as count size tends to infinity

$$f'(\mu_{i,j}) \rightarrow 0 \text{ as } \mu_{i,j} \rightarrow \infty. \tag{4.3}$$

The third property respects the intuition that large counts hold more information on proximity and the distances should reflect this, for example the decrease in distances

between $\mu_{i,j} = 1$ and $\mu_{i,j} = 2$ should be larger than between $\mu_{i,j} = 101$ and $\mu_{i,j} = 102$.

## Exponential transform

The exponential transform used by Makraz (2010) returns distances in the interval $d_{i,j} = (0, 1]$,

$$f(\mu_{i,j}) = \mathrm{e}^{-\alpha\mu_{i,j}} \tag{4.4}$$

where $\alpha > 0$. The parameter $\alpha$ scales the counts and affects the distribution of distances. The upper bound on the $d_{i,j}$ can be motivated as follows: if two megabase intervals are too spatially distant to make contact then $d_{i,j}$ could be any value between 1 and $\infty$. If the $\mathbf{D}$ contains a $d_{i,j} = \infty$ then it cannot be fitted into Euclidean space, also the chromosome is bounded within a territory which should limit the distances between the megabase intervals. The inverse exponential transform used to transform distances into counts is

$$f^{-1}(d_{i,j}) = -\frac{1}{\alpha}\log(d_{i,j}), \tag{4.5}$$

which returns counts in the interval $\mu_{i,j} = (0, \infty]$ for $0 \leq d_{i,j} \leq 1$. Should the fitted configuration $\hat{\mathbf{X}}$ have an interpoint distance $\hat{d}_{i,j} > 1$ then a fitted count of $\hat{\mu}_{i,j} = 0$ is used to avoid negative fitted counts.

## Power transform

The power transform reflects the relationship between $d_{i,j}$ and genomic distance proposed by Mateos-Langerak et al. (2009); the relationship between interaction probability and genomic distance proposed by Lieberman-Aiden et al. (2009) and is a simpler version of the transform used by Hu et al. (2013). The power transform returns distances in the

interval $d_{i,j} = (0, \infty]$,

$$f(\mu_{i,j}) = (b_0 \mu_{i,j})^{\beta} \tag{4.6}$$

where $b_0 > 0$ and $\beta < 0$. The parameter $b_0$ has no effect on the shape of $\hat{\mathbf{X}}$ and we set it at $b_0 = 1$. The parameter $\beta$ determines the relationship between counts and distances and affects the distribution of distances. When using the power transform, an $m_{i,j} = 0$ is transformed into a $d_{i,j} = \infty$, resulting in a $\mathbf{D}$ which cannot be fitted into Euclidean space. To avoid this all $m_{i,j} = 0$ in $\mathbf{M}$ are replaced with $m_{i,j} = 1$, this is called the minimum count adjustment and is denoted by $m_{\text{min}} = 1$. The minimum count adjustment can be increase to any required size $m_{\text{min}} = a$, where all $m_{i,j} < a$ in $\mathbf{M}$ are replaced with $m_{i,j} = a$, although as larger $a$ are used structural information will gradually be erased from $\mathbf{M}$. The inverse power transform is

$$f^{-1}(d_{i,j}) = \frac{d_{i,j}^{\frac{1}{\beta}}}{b_0}, \tag{4.7}$$

which returns counts in the interval $\mu_{i,j} = (0, \infty]$ for $d_{i,j} \geq 0$.

### 4.1.2 Measures of fit

Since $\mathbf{X}$ is unknown, measures of fit are employed to assess $\hat{\mathbf{X}}$. These are divided into score functions which help locate the best fitting configuration, and auxiliary measures which are used to assess the fit.

#### Score functions

Score functions identify the best fitting configuration $\hat{\mathbf{X}}$ for the data. Metric MDS uses the sum of the Pearsons residuals $\chi^2$ and non-metric MDS uses the stress of fit $S_3(\hat{\mathbf{X}})$ (2.14).

**Metric MDS**

Metric MDS is first used to recover the three-dimensional fitted configuration $\hat{\mathbf{X}}$ from the estimated distance matrix $\tilde{\mathbf{D}}$. The fitted distances $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ are first extracted from $\hat{\mathbf{X}}$ using (2.2) with $p = 3$. The $\hat{\mathbf{D}}$ can then be used to provide distance based score functions, such as the stress of fit $S_p(\hat{\mathbf{X}})$ (2.14) used by non-metric MDS. The $\hat{\mathbf{D}}$ can also be used to recover the fitted counts $\hat{\mathbf{U}} = (\hat{\mu}_{i,j})$ by using the inverse transform functions (4.5) or (4.7). The $\hat{\mathbf{U}}$ allows the use of count based score functions. The sum of the squared residuals

$$\text{SSR}(\mathbf{M}, \hat{\mathbf{U}}) = \sum_{i<j} (m_{i,j} - \hat{\mu}_{i,j})^2 \tag{4.8}$$

can be found, which measures the total squared difference between $\mathbf{M}$ and $\hat{\mathbf{U}}$. In practice, $\text{SSR}(\mathbf{M}, \hat{\mathbf{U}})$ can be sensitive to small changes in distances causing large changes in counts. To counter this, residuals $(m_{i,j} - \hat{\mu}_{i,j})$ in (4.8) are divided by $\hat{\mu}_{i,j}$ to give the sum of the Pearsons residuals

$$\chi^2 = \sum_{i<j} \frac{(m_{i,j} - \hat{\mu}_{i,j})^2}{\hat{\mu}_{i,j}}. \tag{4.9}$$

**Non-metric MDS**

Since scale is not preserved in $\hat{\mathbf{X}}$ the $\hat{\mathbf{U}}$ cannot be recovered. This limits the non-metric MDS to the stress of fit statistic $S_p(\hat{\mathbf{X}})$ (2.14), where the $p$ denotes the number of dimensions $\hat{\mathbf{X}}$ is fitted into.

**Auxiliary measures**

In addition to the score functions, auxiliary measures prove useful in assessing $\hat{\mathbf{X}}$.

The relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$ and the distribution of $\tilde{\mathbf{D}}$ provide insight into properties of the transform function. Using a histogram of $\tilde{\mathbf{D}}$ can provide insight into $\hat{\mathbf{X}}$: if the distribution of $\tilde{d}_{i,j}$ are skewed towards large $\tilde{d}_{i,j}$ then horseshoe shaped fitted configurations become more likely. A histogram of $\tilde{\mathbf{D}}$ skewed towards the smaller $\tilde{d}_{i,j}$ is preferable, as smaller $\tilde{d}_{i,j}$ from large $m_{i,j}$ are more accurate.

When using metric MDS plots of the $\hat{\lambda}_k$ can provide insight into the magnitude of noise in $\tilde{\mathbf{D}}$ and dimensionality of $\hat{\mathbf{X}}$. The sizes of $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ relative to $\hat{\lambda}_k$ for $k = 4, \ldots, n$ describe the quantity of information captured in the first three dimensions and how much is distributed into spurious dimensions. The size of the negative $\hat{\lambda}_k$ relative to the positive $\hat{\lambda}_k$ gives insight how much the Euclidean properties of $\tilde{\mathbf{D}}$ are violated. The magnitude criterion (Sibson, 1979) is helpful for differentiating the genuine eigenvalues from the spurious eigenvalues. In the magnitude criterion any positive eigenvalues which are smaller than the absolute magnitude of the lowest negative eigenvalue, can be regarded as spurious eigenvalues.

Visual inspection of $\hat{\mathbf{X}}$ can be used to assess if MDS has fitted $\hat{\mathbf{X}}$ at a local or global minimum. Indications that MDS has found a local minimum might be that the majority of points are clustered together with a select few points at some distance away from the cluster, or that the points fall into layered flat clusters. Visual inspection might also help identify if the horseshoe effect (Section 2.4) has influenced $\hat{\mathbf{X}}$.

Plots of the $\tilde{d}_{i,j}$ against $\hat{d}_{i,j}$ or the $m_{i,j}$ against $\hat{\mu}_{i,j}$ are known as Shepards plots (Cox and Cox (2000) pages 72-73), these are useful to interpret how the MDS fitting moves the distances or counts. In an ideal case, the points in a Shepards plot should be linearly related so a straight line can be plotted through them. In metric MDS, points should line up along the line of zero intercept and gradient one (identity line). Adding a line of best fit with zero intercept to the Shepards plots can help summarize the movement of distances or counts.

Heatmaps are used to visually compare $\tilde{\mathbf{D}}$ with $\hat{\mathbf{D}}$ or $\mathbf{M}$ with $\hat{\mathbf{U}}$. Patterns in the heatmaps of $\tilde{\mathbf{D}}$ and $\mathbf{M}$ relate to structural features of $\mathbf{X}$, and changes in these patterns after fitting can indicate if features are lost or spurious features gained.

### 4.1.3 Fitting algorithm

Fitting $\hat{\mathbf{X}}$ involves scanning across the different values for $\alpha$ when using (4.4) or $\beta$ when using (4.6), until a $\tilde{\mathbf{D}}$ can be found which when fitted into Euclidean space gives a $\hat{\mathbf{X}}$, which minimizes either $\chi^2$ (4.9) or $S_p(\hat{\mathbf{X}})$ (2.14). The process of finding the parameters $\alpha$ or $\beta$, which minimize $\chi^2$ (4.9) or $S_p(\hat{\mathbf{X}})$ (2.14) are outlined as follows.

1. Choose an interval within which we presume $\alpha$ or $\beta$ to lie: $\alpha \in (0, a]$ or $\beta \in (0, b]$, where $a$ or $b$ are chosen such that $\tilde{d}_{i,j} \approx 0$ for $i, j = 1, \ldots, n$.

2. Scan across the interval repeatedly producing $\tilde{\mathbf{D}}$ from $\mathbf{M}$ with either the exponential transform (4.4) or the power transform (4.6). Fitting the $\tilde{\mathbf{D}}$ into $p$ dimensional Euclidean space ($p = 1, 2$ or 3) with either metric or non-metric MDS, to obtain $\hat{\mathbf{X}}$ and calculating $\chi^2$ (4.9) or $S_p(\hat{\mathbf{X}})$ (2.14) recording the values of the score functions.

3. Identify the parameters $\hat{\alpha}$ or $\hat{\beta}$ from the above stage which produces the minimum $\chi^2$ (4.9) or $S_p(\hat{\mathbf{X}})$ (2.14). Then use these values of $\hat{\alpha}$ or $\hat{\beta}$ in the transform function (4.4) or (4.6) to produce $\tilde{\mathbf{D}}$, and fit $\tilde{\mathbf{D}}$ into $p$ dimensional Euclidean space using metric or non-metric MDS to obtain the configuration of best fit $\hat{\mathbf{X}}$.

## 4.2 Estimated chromosome configuration

This section applies the transform functions, MDS and the fitting algorithm to find $\hat{\mathbf{X}}$. Combining the two transform functions and metric or non-metric MDS gives four routes

to recover $\hat{\mathbf{X}}$. These routes are split by MDS method so results between transform functions can be compared. To help identify which transforms function and MDS method is used subscript notation is added to all matrices involved, $E$ or $P$ added for exponential or power transform and $M$ or $NM$ for metric or non-metric MDS respectively; for example $\hat{\mathbf{X}}_{E,M}$ is the estimated chromosome configuration found using the exponential transform (4.4) with metric MDS, and $\tilde{\mathbf{D}}_{E,M}$ is the estimated distance matrix which was obtain by transforming the observed counts $\mathbf{M}$ into distances with the exponential transform and is used by metric MDS to recover $\hat{\mathbf{X}}_{E,M}$.

The count matrix $\mathbf{M} = (m_{i,j})$ from human Chromosome 14's found using Hi-C (Lieberman-Aiden et al., 2009) is used as a trial. It is a medium size chromosome (87 megabases long after removing megabase intervals with poor mapability such as centromeres and telomeres), allowing any insights gained from fitting $\mathbf{M}$ into Euclidean space, to be applicable to small and large chromosomes. Using a small chromosome as a trial might highlight issues in the transforming and fitting process which are not directly applicable to larger chromosomes, the same might be true when using larger chromosomes as a trial.

When using $\mathbf{M}$ from Chromosome 14 with the power transform, only a small fraction of the count matrix elements require adjusting for the minimum count adjustment. Elements of $\mathbf{M}$ with counts of zero which require adjusting for the $m_{\text{min}} = 1$ adjustment, are $m_{14,87}$ (and $m_{87,14}$). Elements of $\mathbf{M}$ with counts of zero or one which require adjusting for the $m_{\text{min}} = 2$ adjustment, are $m_{14,87}$ (and $m_{87,14}$) and $m_{9,45}, m_{22,84}, m_{22,87}, m_{64,87}$ (and $m_{45,9}, m_{84,22}, m_{87,22}, m_{87,64}$). The $m_{\text{min}} = 1$ adjustment adjusts 0.0264 % of the elements of $\mathbf{M}$ while the $m_{\text{min}} = 2$ adjustment adjusts 0.1321 % of the elements of $\mathbf{M}$.

In Chromosome 14 the centromere is removed from the analysis, be deleting the rows and columns in $\mathbf{M}$ corresponding to the centromere, and deleting the row and column of the megabases abutting the centromere. This leaves 87 megabase intervals to fit into space, which are not part of the centromere or sit next to it; this is the procedure used where the

centromere sits at the start of the chromosome. In chromosomes where the centromere sits in the middle of the chromosome, the rows and columns corresponding to the megabase intervals of the centromere are deleted from $\mathbf{M}$, and the rows and columns of the abutting megabase intervals are deleted from $\mathbf{M}$. The abutting megabase intervals are removed as a precaution, should the neighbouring centromere effect their ability to make contact. As the centromere is completely removed from $\mathbf{M}$ it will have no effect on the points in the estimated chromosome configuration. Table 4.1 gives the location of the centromere for each chromosome.

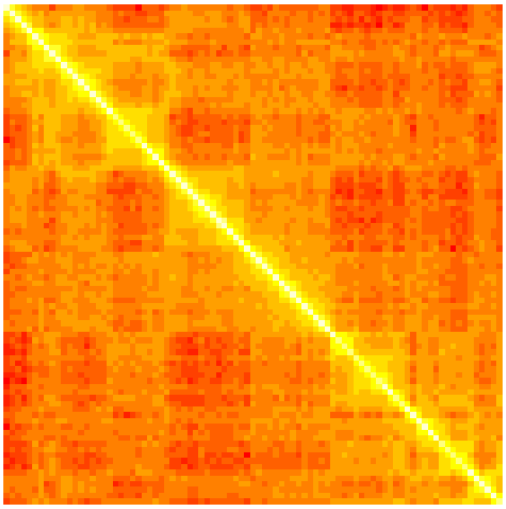
The estimated chromosomes configurations in Appendix Section G have the location of the centromere marked on by a blue line —, the estimated chromosome configurations with this line missing will have the centromere located at the start of the chromosome.

### 4.2.1 Estimated chromosome configuration from metric MDS

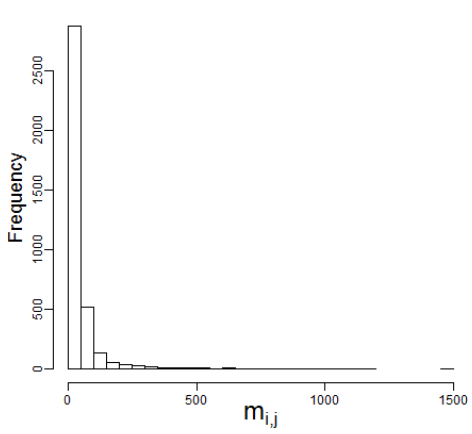
Table 4.2 gives parameter values $\hat{\alpha}$ or $\hat{\beta}$ which minimize $\chi^2$ (4.9), with the corresponding $\text{SSR}(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) and $S_3(\hat{\mathbf{X}})$ (2.14) values. Tables A.1, A.2 and A.3 (Appendix Section A.1) give the parameter and score function values for each chromosomes individually.

The power transform results in Table 4.2 show that the fit corresponding $m_{\text{min}} = 2$ produces a smaller $\chi^2$ (4.9), so this result will be used as the best fitting result for the power transform and will be compared with the exponential transform results. Comparing the results from the two transforms in Table 4.2, the exponential transform produces the smallest $\chi^2$ and accompanying measures of fit, suggesting $\hat{\mathbf{X}}_{E,M}$ is the better estimate of $\mathbf{X}$.

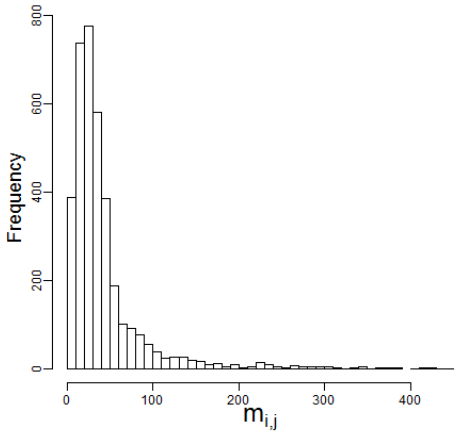
In Figure 4.2a the relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$ (using the exponential transform) appears to tend to zero quickly, with $m_{i,j} < 200$ producing large $\tilde{d}_{i,j}$, $200 \leq m_{i,j} < 600$ producing medium $\tilde{d}_{i,j}$ and $m_{i,j} \geq 600$ producing $\tilde{d}_{i,j} \approx 0$, this relationship could force



(a) Heatmap of the observed counts.




(b) Full histogram.




(c) Truncated histogram.

Figure 4.1: Top panel: heatmap of Chromosome 14's Hi-C count matrix. The bright yellow in the heatmap denote large counts, and as colour moves to darker red count size decreases. Bottom left panel: full histogram of the elements from the lower triangle of Chromosome 14's Hi-C count matrix $\mathbf{M} = (m_{i,j})$. Bottom right panel: truncated histogram to include all $m_{i,j} \leq 450$ from the lower triangle of $\mathbf{M}$.

| Chromosome | Length Mb | Centromere location | Total off diagonal counts | Total diagonal counts |
|---|---|---|---|---|
| 1 | 226 | 123-124 | 842012 | 736117 |
| 2 | 238 | 93-94 | 835841 | 815125 |
| 3 | 194 | 92-93 | 686957 | 612265 |
| 4 | 187 | 51-52 | 577626 | 532770 |
| 5 | 176 | 48-49 | 597039 | 530822 |
| 6 | 166 | 60-61 | 584960 | 487834 |
| 7 | 154 | 60-61 | 521111 | 452522 |
| 8 | 142 | 45-46 | 490413 | 425553 |
| 9 | 121 | 49-50 | 404428 | 338625 |
| 10 | 132 | 41-42 | 474448 | 400507 |
| 11 | 130 | 53-54 | 477916 | 366903 |
| 12 | 129 | 36-37 | 465746 | 344414 |
| 13 | 96 | 0-1 | 3012629 | 237645 |
| 14 | 87 | 0-1 | 316079 | 212533 |
| 15 | 81 | 0-1 | 308043 | 206365 |
| 16 | 78 | 37-38 | 298194 | 215765 |
| 17 | 78 | 0-1 | 317288 | 182448 |
| 18 | 76 | 0-1 | 257251 | 192580 |
| 19 | 54 | 26-27 | 215135 | 121924 |
| 20 | 59 | 28-29 | 244615 | 154790 |
| 21 | 32 | 0-1 | 113223 | 60394 |
| 22 | 34 | 0-1 | 145433 | 77180 |
| X | 150 | 60-61 | 422117 | 419063 |

Table 4.1: Table summarizing the data in the chromosome count matrix. Column one gives the chromosome number. Column two gives the length of the chromosome in megabase intervals, after trimming megabases of poor mapability (centromere region and abutting megabase intervals). The length of the chromosome is also the number of rows and columns in the chromosome count matrix $\mathbf{M}$. Column three gives the megbases intervals the centromere should lie between, 0-1 indicates the centromere is found at the start of the chromosome. Column four and five give the total diagonal counts in $\mathbf{M}$ and total off-diagonal (lower triangle) counts in $\mathbf{M}$.

adjacent megabase intervals to share $d_{i,j} \approx 0$. The histogram in Figure 4.2b displays a large quantity of $\tilde{d}_{i,j} \geq 0.6$, large $\tilde{d}_{i,j}$ over influence $\hat{\mathbf{X}}_{E,M}$ and could force it to take aspects of the horseshoe effect. The clustering of $\tilde{d}_{i,j}$ around a constant could cause

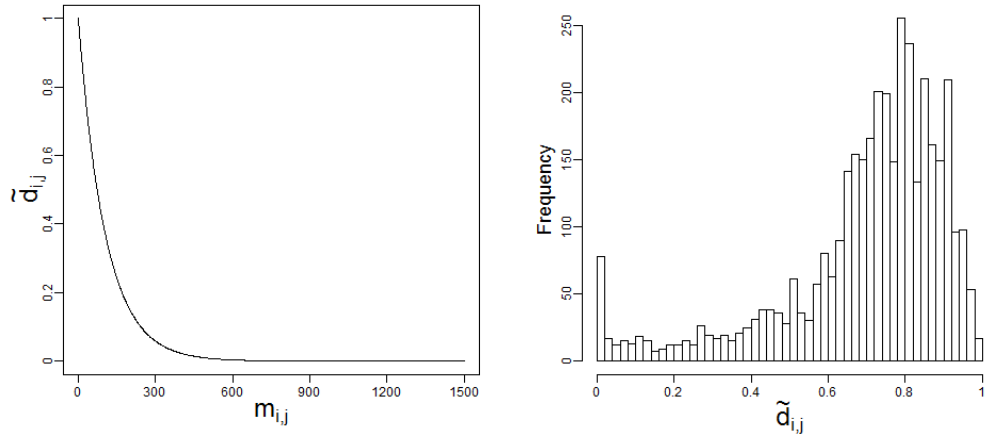
| Transform function | Parmeter | Parameter estimate | $\chi^2$ | SSR$(\mathbf{M}, \hat{\mathbf{U}})$ | $S_3(\hat{\mathbf{X}})$ |
|---|---|---|---|---|---|
| Exponential transform | $\hat{\alpha}$ | 0.0095 | 147172 | $2.7044\times10^7$ | 16.7813% |
| Power transform | $m_{\text{min}} = 1$:$\hat{\beta}$ | -0.4497 | 1171886 | $2.9535\times10^{10}$ | 32.0391% |
| | $m_{\text{min}} = 2$:$\hat{\beta}$ | -0.4796 | 485658 | $7.4257\times10^8$ | 23.73281% |

Table 4.2: Score function data from using metric MDS to obtain an estimated chromosome configuration for Chromosome 14. Column one and two state which transform function has been used and which parameter has been estimated; when using the power transform (4.6) the row is subdivided according which minimum count adjustment has been used. Column three and four give the estimated parameter value and the $\chi^2$ (4.9) value it minimizes. Column five and six give the SSR$(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) and $S_3(\hat{\mathbf{X}})$ (2.14) values found using the estimated parameter values. The $\hat{\alpha}$ for (4.4) and $\hat{\beta}$ for (4.6) are found by applying the fitting algorithm (Section 4.1.3) using the $\chi^2$ score function and fitting into three dimensional Euclidean space with metric MDS, to Chromosome 14's Hi-C count matrix.

problems of indifferentiation (Buja and Swayne, 2002), which in the extreme case where $\tilde{d}_{i,j} = c \,\forall\, i \neq j$ when fitted produces a $n-1$ dimensional simplex where $\hat{d}_{i,j} = c \,\forall\, i \neq j$.

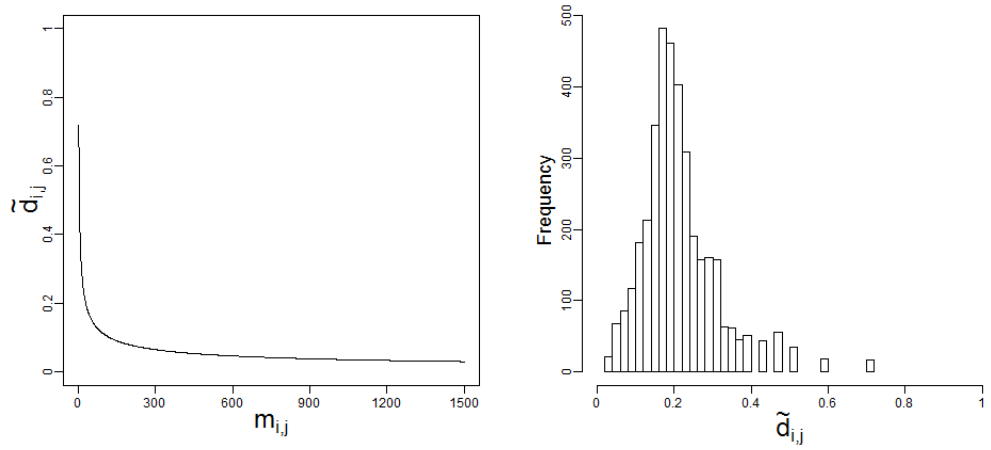
In Figure 4.3a the relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$ (using the power transform) appears to tend to zero gradually, with $m_{i,j} \leq 100$ producing large $\tilde{d}_{i,j}$ and $m > 100$ producing medium to small $\tilde{d}_{i,j}$. The histogram in Figure 4.3b displays a large quantity of $\tilde{d}_{i,j} < 0.4$, reflecting the increased accuracy in smaller distances which are less likely to produce the horseshoe shaped configurations, as discussed in Section 2.4.

Figure 4.4 gives the scree plots of the fitted eigenvalues $\hat{\Lambda}_{E,M}$ or $\hat{\Lambda}_{P,M}$. Ignoring magnitude and inspecting the lead three (genuine) eigenvalues relative to the spurious non-zero eigenvalues, the genuine eigenvalues in $\hat{\mathbf{\Lambda}}_{E,M}$ (Figure 4.4a) appear spaced better from the spurious eigenvalues, than in $\hat{\mathbf{\Lambda}}_{P,M}$. The genuine eigenvalues in $\hat{\mathbf{\Lambda}}_{E,M}$ clearly surpass the largest spurious eigenvalue, whereas for the $\hat{\mathbf{\Lambda}}_{P,M}$ the lead eigenvalue surpasses the second and third eigenvalue which are a similar size to the largest spurious eigenvalue. The negative eigenvalues in $\hat{\mathbf{\Lambda}}_{E,M}$ are small with the absolute magnitude



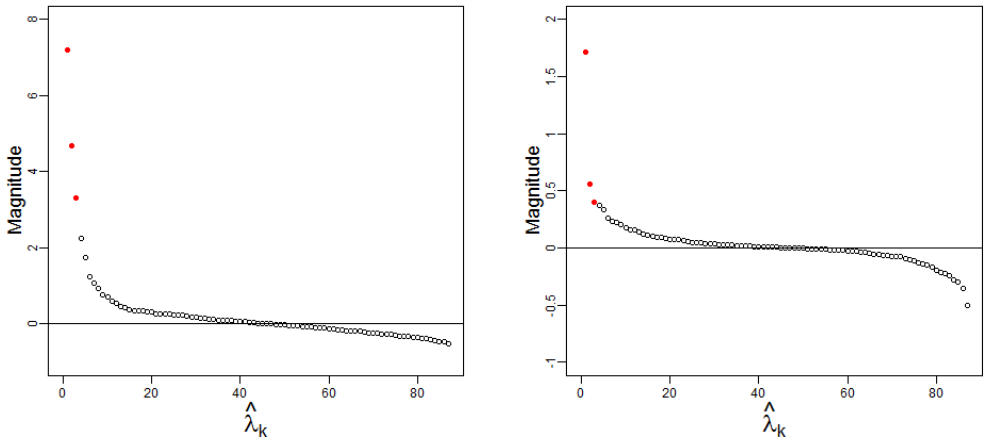

(a) Relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$. (b) Histogram of $\tilde{\mathbf{D}}_{E,M}$.

Figure 4.2: Inspection of the estimated distances $\tilde{\mathbf{D}}_{E,M} = (\tilde{d}_{i,j})$, found from Chromosome 14's Hi-C count matrix $\mathbf{M} = (m_{i,j})$ using the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ (Table 4.2). Left panel: the relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$. Right panel: histogram of the elements in the lower triangle of $\tilde{\mathbf{D}}_{E,M}$.




(a) Relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$. (b) Histogram of $\tilde{\mathbf{D}}_{P,M}$.

Figure 4.3: Inspection of the estimated distances $\tilde{\mathbf{D}}_{P,M} = (\tilde{d}_{i,j})$, found from Chromosome 14's Hi-C count matrix $\mathbf{M} = (m_{i,j})$ using the power transform (4.6) with the $m_{\text{min}} = 2$ adjustment and $\hat{\beta} = -0.4796$ (Table 4.2). Left panel: the relationship between $m_{i,j}$ and $\tilde{d}_{i,j}$. Right panel: histogram of the elements in the lower triangle of $\tilde{\mathbf{D}}_{P,M}$.




(a) Scree plot of $\hat{\mathbf{\Lambda}}_{E,M}$.

(b) Scree plot of $\hat{\mathbf{\Lambda}}_{P,M}$.

Figure 4.4: Left panel: scree plot of the fitted eigenvalues $\hat{\mathbf{\Lambda}}_{E,M}$ (2.7) found from fitting $\tilde{\mathbf{D}}_{E,M}$ into Euclidean space with metric MDS. Right panel: scree plot of the fitted eigenvalues $\hat{\mathbf{\Lambda}}_{P,M}$ (2.7) found from fitting $\tilde{\mathbf{D}}_{P,M}$ into Euclidean space with metric MDS. The three lead fitted eigenvalues are denoted by ●, and the remaining fitted eigenvalues are denoted by ○. The $\tilde{\mathbf{D}}_{E,M}$ are found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ (Table 4.2) to Chromosome 14's Hi-C count matrix $\mathbf{M}$. The $\tilde{\mathbf{D}}_{P,M}$ are found by applying the power transform (4.6) with $m_{\text{min}} = 2$ adjustment and $\hat{\beta} = -0.4796$ (Table 4.2) to $\mathbf{M}$.

of the largest negative eigenvalue smaller than the largest spurious positive eigenvalue, indicating a good Euclidean $\tilde{\mathbf{D}}_{E,M}$. In contrast, the negative eigenvalues in $\hat{\mathbf{\Lambda}}_{P,M}$ (Figure 4.4b) are almost a symmetrical to the positive eigenvalues, with the absolute magnitude of the largest negative eigenvalue similar size to the second eigenvalue, indicating a poor Euclidean $\tilde{\mathbf{D}}_{P,M}$. The proportion of information projected into the first three dimensions tells a similar story, the exponential transform performs better as the spurious eigenvalues influence $\theta_{1:3}$ (2.12) less than in the power transform.

The fitted configuration $\hat{\mathbf{X}}_{E,M}$ in Figure 4.5 appears as a horseshoe when plotted in the first and second dimensions although in the first and third dimensions the cubic polynomial relationship arrangement does not appear to be present. The centre of $\hat{\mathbf{X}}_{E,M}$ appears hollow. Following the path of the chromosome in $\hat{\mathbf{X}}_{E,M}$ the points appear to

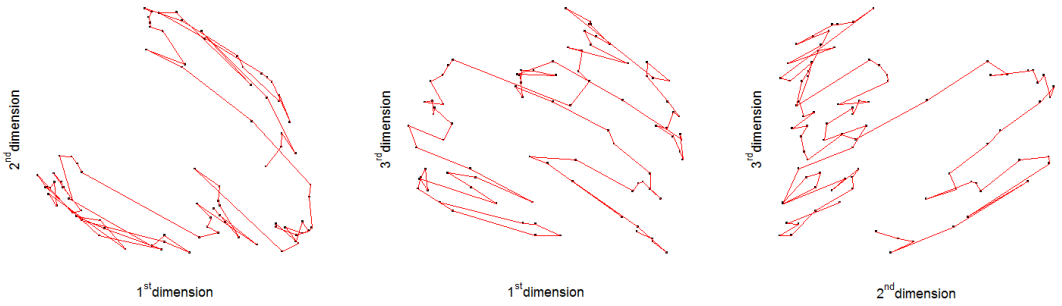



Figure 4.5: Perspectives of Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$. The origin of the megabase interval is denoted by the point . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}_{E,M}$ is found by fitting the estimated distances $\tilde{\mathbf{D}}_{E,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{E,M}$ is found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ (Table 4.2) to Chromosome 14's Hi-C count matrix.

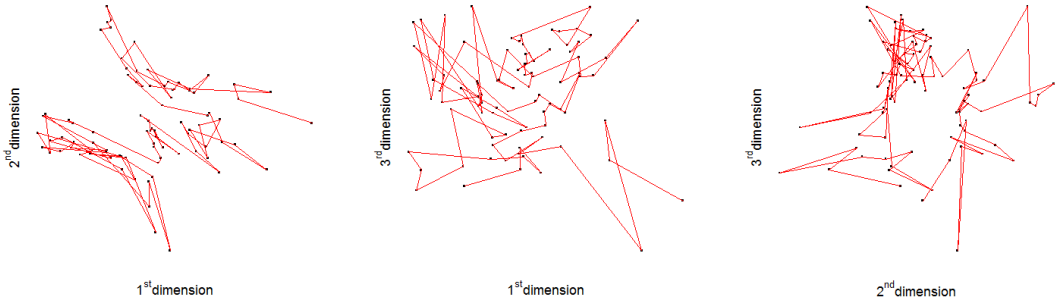



Figure 4.6: Perspectives of Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$. The origin of the megabase interval is denoted by the point . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}_{P,M}$ is found by fitting the estimated distances $\tilde{\mathbf{D}}_{P,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{P,M}$ is found by applying the power transform (4.6) with $m_{\text{min}} = 2$ adjustment and $\hat{\beta} = -0.4796$ (Table 4.2) to Chromosome 14's Hi-C count matrix.

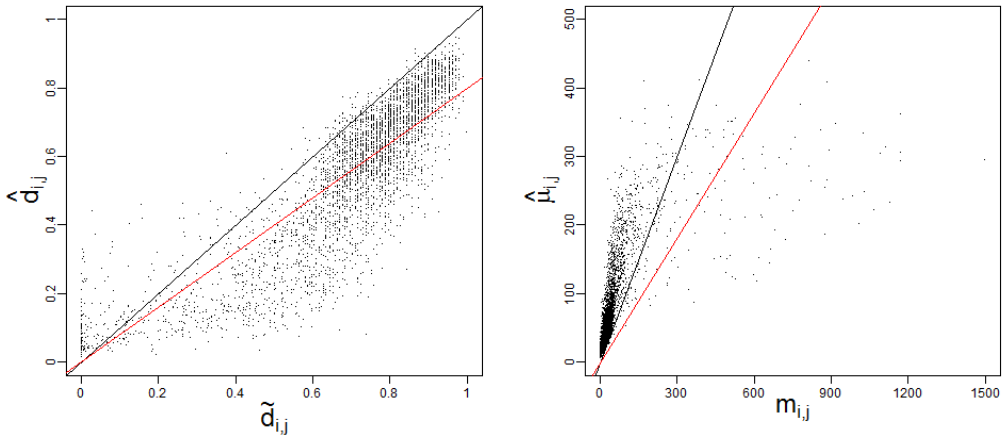
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_{1:3}$ |
|---|---|---|---|---|
| Exponential transform | 17.887% | 11.604% | 8.191% | 37.682% |
| Power transform | 16.570% | 5.356% | 3.875% | 25.801% |

Table 4.3: Percentage of information projected into the first three dimensions $\theta_1$, $\theta_2$ and $\theta_3$ and total percentage of information projected into the first three dimensions $\theta_{1:3}$. The $\theta_1$, $\theta_2, \theta_3$ and $\theta_{1:3}$ values are found by substituting the fitted eigenvalues in $\hat{\mathbf{\Lambda}}_{E,M}$ (for (4.4)) or $\hat{\mathbf{\Lambda}}_{P,M}$ (for (4.6)) into (2.11) and (2.12). The fitted eigenvalues $\hat{\mathbf{\Lambda}}_{E,M}$ (2.7) are found from fitting $\tilde{\mathbf{D}}_{E,M}$ into Euclidean space with metric MDS. The $\tilde{\mathbf{D}}_{E,M}$ are found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ (Table 4.2) to Chromosome 14's Hi-C count matrix $\mathbf{M}$. The fitted eigenvalues $\hat{\mathbf{\Lambda}}_{P,M}$ (2.7) are found from fitting $\tilde{\mathbf{D}}_{P,M}$ into Euclidean space with metric MDS. The $\tilde{\mathbf{D}}_{P,M}$ are found by applying the power transform (4.6) with $m_{\text{min}} = 2$ adjustment and $\hat{\beta} = -0.4796$ (Table 4.2) to $\mathbf{M}$.

meander and form clusters, which could correspond to features of $\mathbf{X}$ captured at a local scale. The fitted configuration $\hat{\mathbf{X}}_{P,M}$ in Figure 4.6 presents a less obvious horseshoe in the first and second dimensions. In the first and third dimensions the polynomial relationship does not appear to be present and the centre of $\hat{\mathbf{X}}_{P,M}$ appears less hollow. The path of the chromosome in $\hat{\mathbf{X}}_{P,M}$ meanders more chaotically but still forms clusters.

In the Shepards plot of distances from the exponential transform in Figure 4.7a, the fitted configuration has filled a void in medium distances by increasing small distances and decreasing large distances, and the gradient of the line of best fit below one suggesting a general decrease in distances. It is more difficult to interpret the Shepards plot of counts from the exponential transform in Figure 4.7b. The majority of counts are placed above the identity line reflecting the decrease in distances, although the line of best fit is below one this could be biased by a larger decrease in large counts.

In the Shepards plot of distances from the power transform in Figure 4.8a, the plotted points lie close to identity line with some variation indicating a good fit, and the gradient of the line of best fit is almost equal to one suggesting a balanced adjustment of distances.

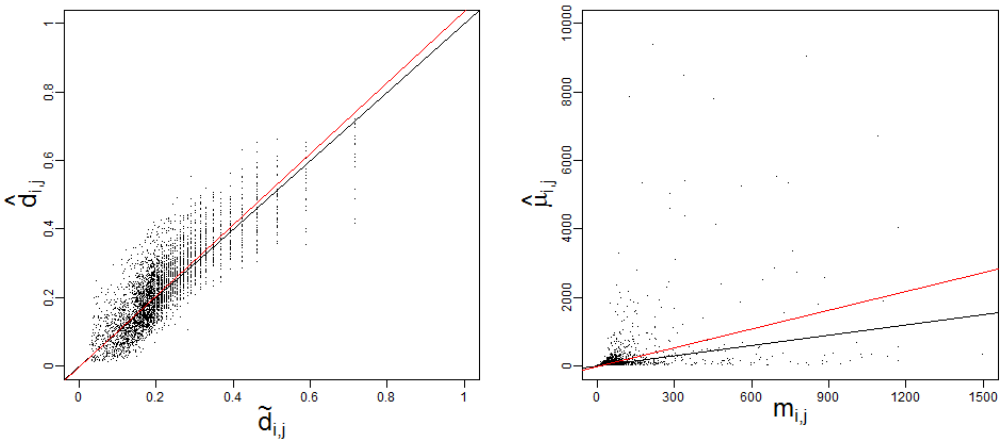



(a) Distances (b) Counts

Figure 4.7: Left panel: Shepards plot of the fitted distances $\hat{\mathbf{D}}_{E,M}$ and estimated distances $\tilde{\mathbf{D}}_{E,M}$. Right panel: Shepards plot of the fitted counts $\hat{\mathbf{U}}_{E,M}$ and Chromosome 14's Hi-C counts $\mathbf{M}$. In both figures the identity line is denoted by ——, and the line of best fit with zero intercept and gradient 0.7980 for distances or 0.6066 for counts is denoted by ——. The elements of $\hat{\mathbf{U}}_{E,M}$ are obtained by inputting the elements of $\hat{\mathbf{D}}_{E,M}$ into the inverse exponential transform (4.5) with $\hat{\alpha} = 0.0095$ (Table 4.2). The elements of $\hat{\mathbf{D}}_{E,M}$ are extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$, using (2.2) with $p = 3$. The estimated configuration $\hat{\mathbf{X}}_{E,M}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{E,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{E,M}$ is found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ to $\mathbf{M}$.

The Shepards plot of counts from the power transform in Figure 4.8b displays a large average increase in count size. This is driven by small distances decreasing and been magnified through (4.7) to produce very large increases in large counts, for example $m_{61,63} = 216$ becomes $\hat{\mu}_{61,63} = 9375$ after fitting. This magnification explains the $\text{SSR}(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) from the power transform being in the order of the tens of billions.

In Figure 4.9a the relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$ for the exponential transform (4.4) returns the majority of fitted counts $\hat{\mu}_{i,j} \leq 200$ and only very small $\hat{d}_{i,j}$ produce $\hat{\mu}_{i,j} > 200$. This can be seen in the histogram in Figure 4.9b with the majority of the fitted counts situated $\hat{\mu}_{i,j} \leq 200$. The histogram in Figure 4.9b is similar in shape to the truncated histogram of the observed counts $m_{i,j}$ Figure 4.1c, showing that the inverse of




(a) Distances (b) Counts

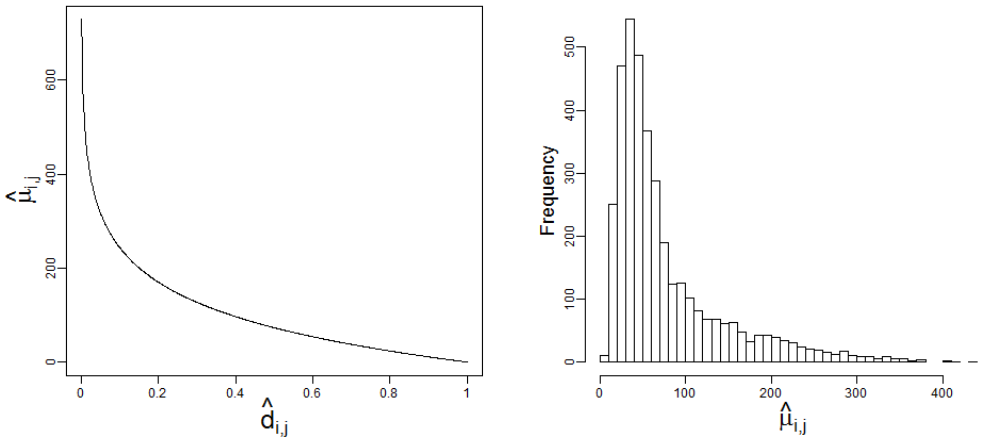
Figure 4.8: Left panel: Shepards plot of the fitted distances $\hat{\mathbf{D}}_{P,M}$ and estimated distances $\tilde{\mathbf{D}}_{P,M}$. Right panel: Shepards plot of the fitted counts $\hat{\mathbf{U}}_{P,M}$ and Chromosome 14's Hi-C counts $\mathbf{M}$. In both figures the identity line is denoted by ——, and the line of best fit with zero intercept and gradient 1.0345 for distances or 1.8128 for counts is denoted by ——. The elements of $\hat{\mathbf{U}}_{P,M}$ are obtained by inputting the elements of $\hat{\mathbf{D}}_{P,M}$ into the inverse power transform (4.7) with $\hat{\beta} = -0.4796$ (Table 4.2). The elements of $\hat{\mathbf{D}}_{P,M}$ are extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$, using (2.2) with $p = 3$. The estimated configuration $\hat{\mathbf{X}}_{P,M}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{P,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{P,M}$ is found by applying the power transform (4.6) with $m_{\text{min}} = 2$ and $\hat{\beta} = -0.4796$ to $\mathbf{M}$.

the exponential transform (4.5) recovers the distribution of small $m_{i,j}$ fairly well.

In Figure 4.10a the relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$ for the power transform (4.6) returns the majority of the counts below $\hat{\mu}_{i,j} \leq 1000$ but very small $\hat{d}_{i,j}$ have the potential to return very large $\hat{\mu}_{i,j}$. In Figure 4.10b the histogram reveals an extremely long tail with the majority of the $\hat{\mu}_{i,j} \leq 100$ and few very large $\hat{\mu}_{i,j}$.

The histogram in Figure 4.10d is similar to the truncated histogram of the observed counts in Figure 4.1c, indicating that the inverse of the power transform (4.7) recovers the distribution of small $m_{i,j}$ well but performs poor for large $m_{i,j}$.

The patterns in the heatmaps in Figure 4.11 are somewhat blurred in $\tilde{\mathbf{D}}_{E,M}$ and $\tilde{\mathbf{D}}_{P,M}$

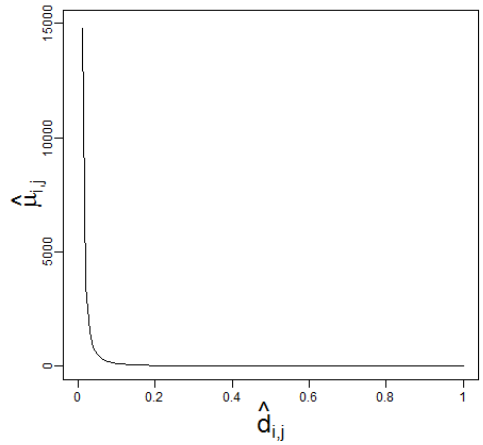



(a) Relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$.

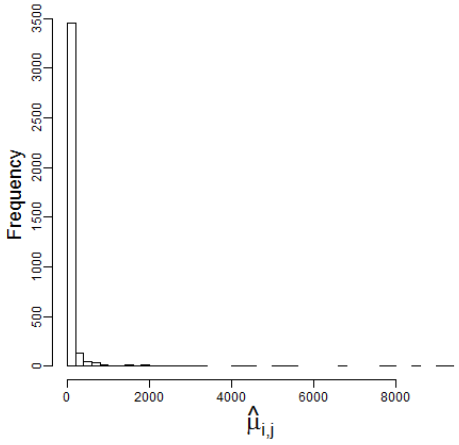
(b) Histogram of $\hat{\mathbf{U}}_{E,M}$

Figure 4.9: Inspection of the fitted counts $\hat{\mathbf{U}}_{E,M} = (\hat{\mu}_{i,j})$, from Chromosome 14's estimated configuration's $\hat{\mathbf{X}}_{E,M}$ fitted distances $\hat{\mathbf{D}}_{E,M} = (\hat{d}_{i,j})$. Left panel: the relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$. Right panel: histogram of the elements in the lower triangle of $\hat{\mathbf{U}}_{E,M}$. The elements of $\hat{\mathbf{U}}_{E,M}$ are obtained by inputting the elements of $\hat{\mathbf{D}}_{E,M}$ into the inverse exponential transform (4.5) with $\hat{\alpha} = 0.0095$ (Table 4.2). The elements of $\hat{\mathbf{D}}_{E,M}$ are extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$, using (2.2) with $p = 3$. The estimated configuration $\hat{\mathbf{X}}_{E,M}$ is found by fitting the estimated distances $\tilde{\mathbf{D}}_{E,M}$ into three dimensional Euclidean space with metric MDS. The $\tilde{\mathbf{D}}_{E,M}$ is found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0095$ to Chromosome 14's Hi-C count matrix.

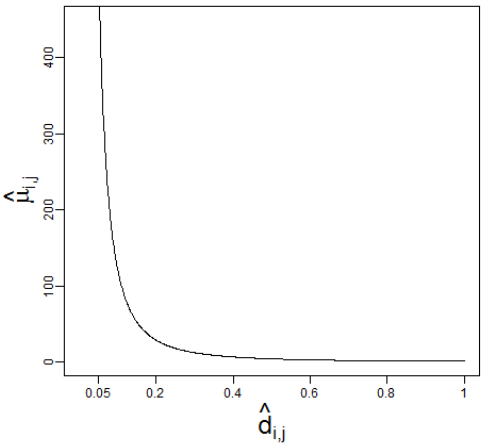
but are sharpened in $\hat{\mathbf{D}}_{E,M}$ and $\hat{\mathbf{D}}_{P,M}$. This suggests the metric MDS removes a lot of the noise from the estimated distance matrices. Three patterns can be distinguished. The first pattern divides the heatmap into nine blocks and subdivides each block into four smaller blocks. The second pattern is the band of small distances running next to the main diagonal, corresponding to the band of large counts running next to the main diagonal of $\mathbf{M}$. The third pattern is additional structure on inside the three blocks found on the diagonal. Combinations of these patterns can help explain the structure in $\hat{\mathbf{X}}$, for example the second pattern reflects local structure in $\hat{\mathbf{X}}$ and combined with the first pattern this reflects clumping of chromatin along the chromosome. Fitting also appears to remove noise, returning cleaner looking heatmaps for the fitted distance matrices $\hat{\mathbf{D}}_{E,M}$



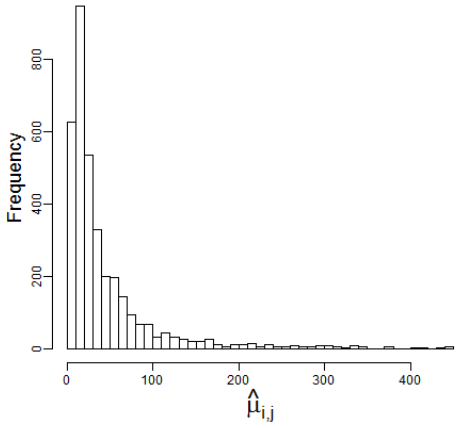

(a) Relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$.




(b) Histogram of $\hat{\mathbf{U}}_{P,M}$.

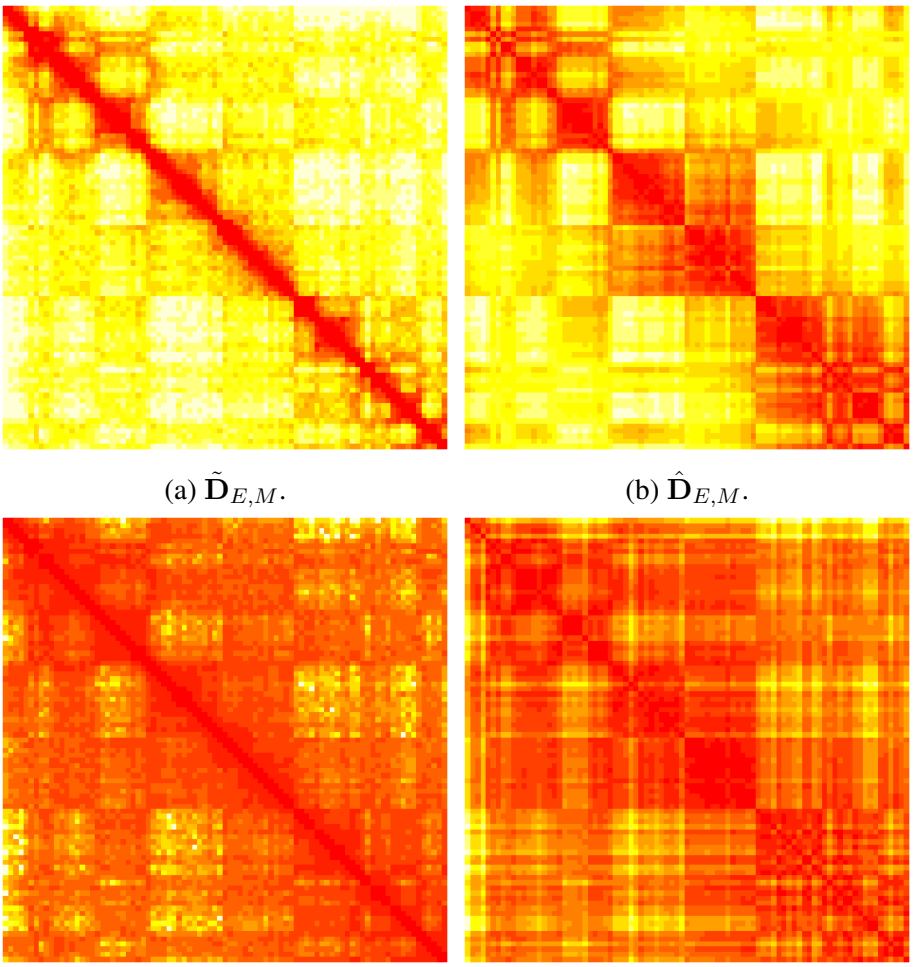



(c) Relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$.




(d) Truncated histogram of $\hat{\mathbf{U}}_{P,M}$.

Figure 4.10: Inspection of the fitted counts $\hat{\mathbf{U}}_{P,M} = (\hat{\mu}_{i,j})$, from Chromosome 14's estimated configuration's $\hat{\mathbf{X}}_{P,M}$ fitted distances $\hat{\mathbf{D}}_{P,M} = (\hat{d}_{i,j})$. Figure 4.10a: the relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$. Figure 4.10b: histogram of the elements in the lower triangle of $\hat{\mathbf{U}}_{P,M}$. Figure 4.10c: the relationship between $\hat{d}_{i,j}$ and $\hat{\mu}_{i,j}$, truncated to include all $\hat{\mu}_{i,j} \leq 450$. Figure 4.10d: histogram of the elements in the lower triangle of $\hat{\mathbf{U}}_{P,M}$, truncated to include all $\hat{\mu}_{i,j} \leq 450$. The elements of $\hat{\mathbf{U}}_{P,M}$ are obtained by inputting the elements of $\hat{\mathbf{D}}_{P,M}$ into the inverse power transform (4.7) with $\hat{\beta} = -0.4796$ (Table 4.2). The elements of $\hat{\mathbf{D}}_{P,M}$ are extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$, using (2.2) with $p = 3$. The estimated configuration $\hat{\mathbf{X}}_{P,M}$ is found by fitting the estimated distances $\tilde{\mathbf{D}}_{P,M}$ into three dimensional Euclidean space with metric MDS. The $\tilde{\mathbf{D}}_{P,M}$ is found by applying the power transform (4.6) with $m_{\text{min}} = 2$ and $\hat{\beta} = -0.4796$ to Chromosome 14's Hi-C count matrix.



(a) $\tilde{\mathbf{D}}_{E,M}$. (b) $\hat{\mathbf{D}}_{E,M}$.

(c) $\tilde{\mathbf{D}}_{P,M}$. (d) $\hat{\mathbf{D}}_{P,M}$.

Figure 4.11: Row one: heatmaps of estimated distance $\tilde{\mathbf{D}}_{E,M}$ and fitted distance $\hat{\mathbf{D}}_{E,M}$ matrices found using the exponential transform (4.4) and metric MDS. Row two: heatmaps of estimated distance $\tilde{\mathbf{D}}_{P,M}$ and fitted distance $\hat{\mathbf{D}}_{P,M}$ matrices found using the power transform (4.6) and metric MDS. In each heatmap distance increases as colour brightens from dark red to bright yellow. The matrix $\hat{\mathbf{D}}_{E,M}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$ using (2.2) with $p = 3$, where $\hat{\mathbf{X}}_{E,M}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{E,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{E,M}$ is found by applying the exponential transform with $\hat{\alpha} = 0.0095$ (Table 4.2) to Chromosome 14's Hi-C count matrix $\mathbf{M}$. The matrix $\hat{\mathbf{D}}_{P,M}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$ using (2.2) with $p = 3$, where $\hat{\mathbf{X}}_{P,M}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{P,M}$ into three dimensional Euclidean space with metric MDS. The matrix $\tilde{\mathbf{D}}_{P,M}$ is found by applying the power transform with $m_{\text{min}} = 2$ and $\hat{\beta} = -0.4796$ (Table 4.2) to $\mathbf{M}$.

| Transform function | Parameter | Parameter estimate | $S_3(\hat{\mathbf{X}})$ |
|---|---|---|---|
| Exponential transform | $\hat{\alpha}$ | 0.0961 | 12.1728% |
| Power function | $\hat{\beta}$:$m_{\text{min}} = 1$ | -0.3040 | 12.1724% |

Table 4.4: Score function data from using non-metric MDS to obtain an estimated chromosome configuration for Chromosome 14. Column one and two state which transform function has been used and which parameter has been estimated. Column three and four give the estimated parameter value and the $S_3(\hat{\mathbf{X}})$ (2.14) value it minimizes. The $\hat{\alpha}$ for (4.4) and $\hat{\beta}$ for (4.6) are found by applying the fitting algorithm (Section 4.1.3) using the $S_3(\hat{\mathbf{X}})$ score function and fitting into three dimensional Euclidean space with non-metric MDS, to Chromosome 14's Hi-C count matrix.

and $\hat{\mathbf{D}}_{P,M}$.

### 4.2.2 Estimated chromosome configuration from non-metric MDS

The fitting algorithm in Section 4.1.3 was applied with non-metric MDS, finding the transform function parameter values $\hat{\alpha}$ for (4.4) or $\hat{\beta}$ for (4.6), which produces the estimated distances matrix $\tilde{\mathbf{D}}$ which minimizes $S_3(\hat{\mathbf{X}})$ (2.14). Non-metric MDS relies on the rank ordering of the distances so little difference is expected between $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ as the rank ordering of the respective estimated distance will not vary between transforms. The $m_{\text{min}} = 1$ adjustment will only be made for the power transform to avoid $\tilde{d}_{i,j} = \infty$; the $m_{\text{min}} = 2$ adjustment is not required as non-metric MDS relies on the rank ordering of the distances. Table 4.4 gives the transform function parameter values $\hat{\alpha}$ (4.4) or $\hat{\beta}$ (4.6) ($m_{\text{min}} = 1$) which minimize $S_3(\hat{\mathbf{X}})$. Tables A.4 A.5 and A.6 (in Appendix Section A.2) give the parameter and $S_3(\hat{\mathbf{X}})$ score function values for all the chromosomes.

$S_3(\hat{\mathbf{X}})$ (2.14) values in Table 4.4 are almost equal suggesting $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ are

equally valid estimates of $\mathbf{X}$. Since rank ordering of distances are used, the small difference in $S_3(\hat{\mathbf{X}})$ can be attributed to the small order change from the $m_{\text{min}} = 1$ adjustment.

The fitted configurations $\hat{\mathbf{X}}_{E,NM}$ (Figure 4.12) and $\hat{\mathbf{X}}_{P,NM}$ (Figure 4.13) now appear identical to each other. In both $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ a horseshoe shape is present when the first and second dimensions are plotted together, although the cubic polynomial shape is not present when the first and third dimensions are plotted together. The points in $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ appear to meander about and form cluster. Both fitted configurations $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ have hollow centres.

The heatmaps in Figure 4.14 share broadly similar patterns to their metric equivalents of blocks, diagonal bands and patterns at the ends of the bands. The fitting also appears to remove noise producing cleaner looking heatmaps for fitted distance matrices $\hat{\mathbf{D}}_{E,NM}$ and $\hat{\mathbf{D}}_{P,NM}$.

### 4.2.3 Discussion

Metric MDS provides more options when measuring fit than non-metric MDS, but suffers in that it requires a Euclidean distance matrix to recover $\hat{\mathbf{X}}$. When fitting with metric MDS, the fitted counts $\hat{\mu}_{i,j}$ appear sensitive to small changes in small distances, which can cause score functions to blow up, distance based score functions such as $S_3(\hat{\mathbf{X}})$ (2.14) can avoid this problem.

Since non-metric MDS operates on the rank ordering of the distances simpler transform functions were tested, (such as $\tilde{d}_{i,j} = c - m_{i,j}$ and $\tilde{d}_{i,i} = 0$, where $c$ is some constant such that $c > m_{i,j} \, \forall \, i \neq j$). When these estimated distance matrices were fitted with non-metric MDS similar looking $\hat{\mathbf{X}}$ were obtained.

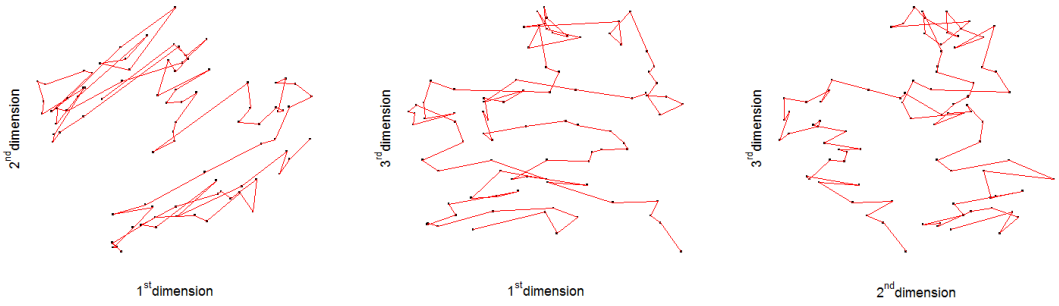



Figure 4.12: Perspectives of Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,NM}$. The origin of the megabase interval is denoted by the point . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}_{E,NM}$ is found by fitting the estimated distance matrix $\tilde{\mathbf{D}}_{E,NM}$ into three dimensional Euclidean space with non-metric MDS. The matrix $\tilde{\mathbf{D}}_{E,NM}$ is found by applying the exponential transform (4.4) with $\hat{\alpha} = 0.0961$ (Table 4.4) to Chromosome 14's Hi-C count matrix.

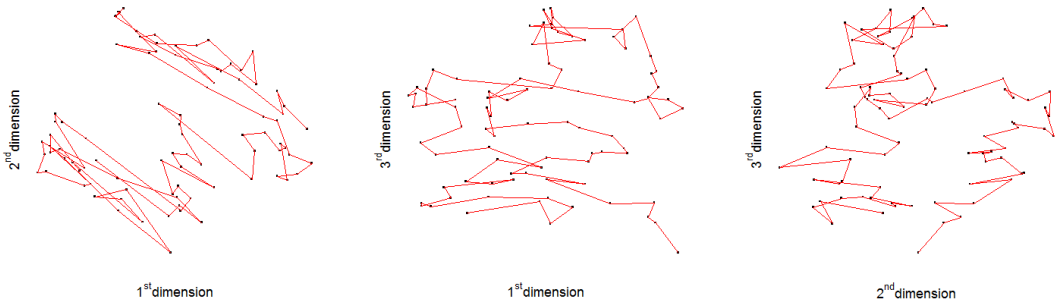
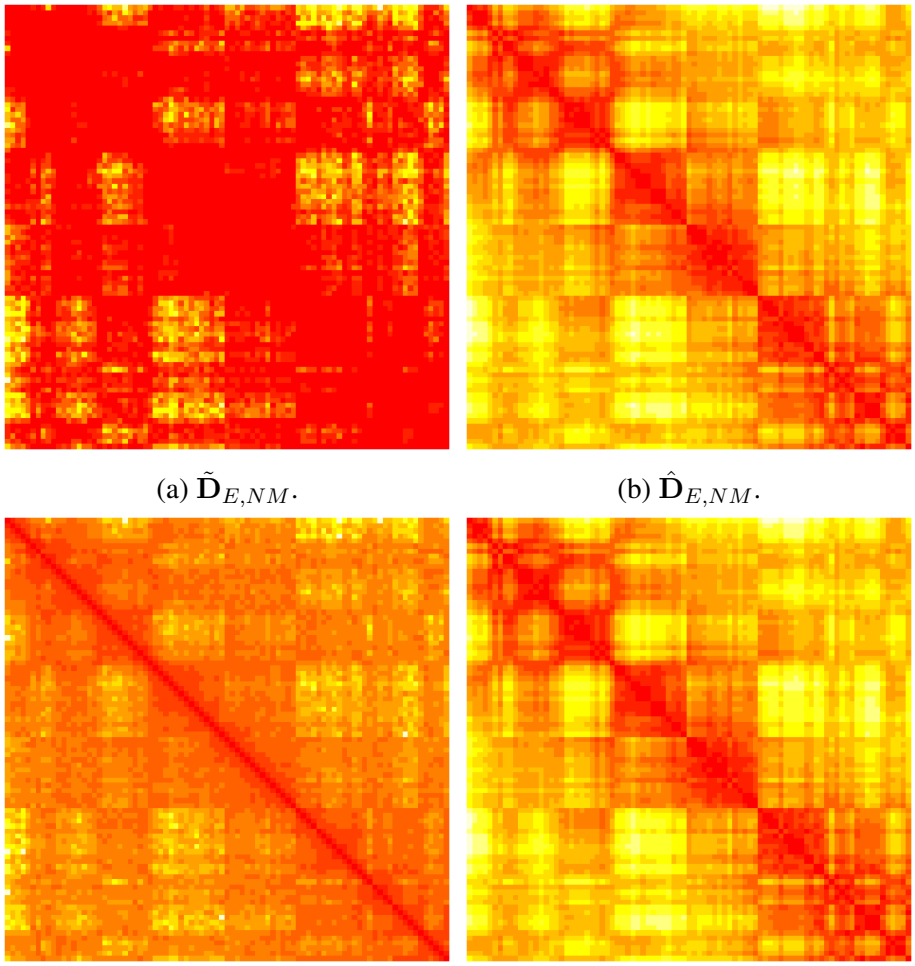



Figure 4.13: Perspectives of Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,NM}$. The origin of the megabase interval is denoted by the point . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}_{P,NM}$ is found by fitting the estimated distance matrix $\tilde{\mathbf{D}}_{P,NM}$ into three dimensional Euclidean space with non-metric MDS. The matrix $\tilde{\mathbf{D}}_{P,NM}$ is found by applying the power transform (4.6) with $m_{\text{min}} = 1$ adjustment and $\hat{\beta} = -0.3040$ (Table 4.4) to Chromosome 14's Hi-C count matrix.



(a) $\tilde{\mathbf{D}}_{E,NM}$. (b) $\hat{\mathbf{D}}_{E,NM}$.

(c) $\tilde{\mathbf{D}}_{P,NM}$. (d) $\hat{\mathbf{D}}_{P,NM}$.

Figure 4.14: Row one: heatmaps of estimated distance $\tilde{\mathbf{D}}_{E,NM}$ and fitted distance $\hat{\mathbf{D}}_{E,NM}$ matrices found using the exponential transform (4.4) and non-metric MDS. Row two: heatmaps of estimated distance $\tilde{\mathbf{D}}_{P,NM}$ and fitted distance $\hat{\mathbf{D}}_{P,NM}$ matrices found using the power transform (4.6) and non-metric MDS. In each heatmap distance increases as colour brightens from dark red to bright yellow. The matrix $\hat{\mathbf{D}}_{E,NM}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,NM}$ using (2.2) with $p = 3$, where $\hat{\mathbf{X}}_{E,NM}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{E,NM}$ into three dimensional Euclidean space with non-metric MDS. The matrix $\tilde{\mathbf{D}}_{E,NM}$ is found by applying the exponential transform with $\hat{\alpha} = 0.0961$ (Table 4.4) to Chromosome 14's Hi-C count matrix $\mathbf{M}$. The matrix $\hat{\mathbf{D}}_{P,NM}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,NM}$, using (2.2) with $p = 3$, where $\hat{\mathbf{X}}_{P,NM}$ is found by fitting the matrix $\tilde{\mathbf{D}}_{P,NM}$ into three dimensional Euclidean space with non-metric MDS. The matrix $\tilde{\mathbf{D}}_{P,NM}$ is found by applying the power transform with $m_{\text{min}} = 1$ and $\hat{\beta} = -0.3040$ (Table 4.4) to $\mathbf{M}$.

## 4.3 Comparing estimated chromosome configurations

The combination of two transform functions and two MDS methods provides four estimated chromosome configurations $\hat{\mathbf{X}}$'s. To measure the difference between the $\hat{\mathbf{X}}$'s, two measures of shape difference were used, one based on the Procustes sum of squares and another based on the fitted distance matrices (Segal et al., 2014). Subscript notation from the previous section will be retained (E=exponential, P=power, M=metric and NM=non-metric).

### 4.3.1 Measuring the difference between estimated chromosome configurations

#### Prescaled Procrustes distance

The prescaled Procrustes distance $\text{POSS}(\mathbf{X}, \mathbf{Y})$ is a modification on the Procrustes distance $\text{OSS}(\mathbf{X}, \mathbf{Y})$ (2.16). The prescaled Procrustes distance is

$$\text{POSS}(\mathbf{X}, \mathbf{Y}) = \text{OSS}\left(\frac{\mathbf{Y}}{||\mathbf{Y}||}, \frac{\mathbf{X}}{||\mathbf{X}||}\right) \tag{4.10}$$

where $||\mathbf{Y}|| = (\text{tr}(\mathbf{Y}\mathbf{Y}^T))^{\frac{1}{2}}$. This is one measure of shape difference between configurations. Prescaling the configurations prevents issues with scaling within the Procrustes distance such as $\text{OSS}(\mathbf{Y}, \mathbf{X}) \neq \text{OSS}(\mathbf{X}, \mathbf{Y})$.

#### Distance differencing

Difference differencing $\text{DD}(\mathbf{X}, \mathbf{Y})$ calculates the distances between two fitted distance matrices (2.2), from the fitted configurations. Let $\mathbf{D}_Y = (d_{y,i,j})$ be the distance matrix

for configuration $\mathbf{Y}$, and $\mathbf{D}_x = (d_{x,i,j})$ be the distance matrix for configuration $\mathbf{X}$. The distances are scaled

$$d^*_{y,i,j} = \frac{d_{y,i,j}}{\sum_{i<j} d_{y,i,j}},$$

then distance is calculated between $\mathbf{D}^*_Y$ and $\mathbf{D}^*_X$ by

$$\text{DD}(\mathbf{X}, \mathbf{Y}) = \left( \sum_{i<j} (d^*_{y,i,j} - d^*_{x,i,j})^2 \right)^{\frac{1}{2}}. \tag{4.11}$$

### 4.3.2 Difference between the estimated chromosome configurations

| | | | |
|---|---|---|---|
| $\hat{\mathbf{X}}_{E,M}$ | 0.2063 | 0.0409 | 0.0411 |
| 0.0080 | $\hat{\mathbf{X}}_{P,M}$ | 0.1441 | 0.1416 |
| 0.0033 | 0.0064 | $\hat{\mathbf{X}}_{E,NM}$ | 0.0002 |
| 0.0033 | 0.0063 | 0.0002 | $\hat{\mathbf{X}}_{P,NM}$ |

Table 4.5: Matrix of shape difference measures between Chromosome 14's estimated configurations (Section 4.2). Upper triangle: POSS$(\mathbf{X}, \mathbf{Y})$ (4.10) values. Lower triangle: DD$(\mathbf{X}, \mathbf{Y})$ (4.11) values. Diagonal entries denote which configurations the row or column values refer to.

From Table 4.5 $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ are almost identical in shape according to both measures, with $\hat{\mathbf{X}}_{E,M}$ very close in shape to these two. The configuration $\hat{\mathbf{X}}_{P,M}$ appears to be dissimilar from the other $\hat{\mathbf{X}}$'s, co-producing the largest distance measures. The $\hat{\mathbf{X}}$'s can be sorted into two groups of configurations, group one containing $\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$ and group two only containing $\hat{\mathbf{X}}_{P,NM}$.

### 4.3.3 All chromosome configurations

Applying the four routes to estimate chromosome configuration gives four estimated chromosome configurations for each chromosome. Then by measuring the prescaled Procrustes distances (4.10) between the four configurations, and applying average linkage cluster analysis, the four configurations can be sorted into two groups of configurations. Applying this to Chromosome 14 we found group one contained $\hat{\mathbf{X}}_{E,M}, \hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$, while group two contained $\hat{\mathbf{X}}_{P,M}$, where $E$ or $P$ denoted if the exponential transform (4.4) or power transform (4.6) was used and $M$ or $NM$ denoted if metric or non-metric MDS was used. Applying this method to each chromosome we found two thirds of the chromosomes group like this, and one third of the chromosomes group differently. In Appendix Section G gives the figures of the two estimated configurations for each chromosome, where one figure represents the estimated configuration from group one and the other figure from group two.

## 4.4 Cluster analysis

The heatmaps of the estimated distance matrices $\tilde{\mathbf{D}}$ and fitted distance matrices $\hat{\mathbf{D}}$ (see Figures 4.11 and 4.14) share a plaid pattern, which divides the matrices into nine blocks and subdivides the blocks into four smaller blocks. These blocks could represent discrete units of structure within the Chromosome 14. The three blocks located on the diagonal of the distance matrices contain the distances between points within the units, and the six blocks located off the diagonal contain the distance between the points in different units.

### 4.4.1 Application of cluster analysis

In Figures 4.15a and 4.15b, the blocks in $\hat{\mathbf{D}}_{E,M}$ and $\hat{\mathbf{D}}_{P,M}$ were investigated using average linkage clustering (Everitt et al. (2001) pages 55-89). The cluster analysis was combined

| Chromosome | Group one | Group two |
|---|---|---|
| 1 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 2 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 3 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 4 | $(P, M), (E, NM), (P, NM)$ | $(E, M)$ |
| 5 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 6 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 7 | $(P, M), (E, M), (P, NM)$ | $(E, NM)$ |
| 8 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 9 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 10 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 11 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 12 | $(P, M), (E, M), (P, NM)$ | $(E, NM)$ |
| 13 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 14 | $(E, M), (E, NM), (P, NM)$ | $(P, M)$ |
| 15 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 16 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 17 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 18 | $(E, NM), (P, M)$ | $(P, MM), (E, M)$ |
| 19 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 20 | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |
| 21 | $(E, M), (P, NM)$ | $(P, M), (E, NM)$ |
| 22 | $(P, M), (E, M), (P, NM)$ | $(E, NM)$ |
| X | $(P, M)$ | $(E, M), (E, NM), (P, NM)$ |

Table 4.6: Table detailing how the four estimated configuration for each chromosome are grouped together. The prescaled Procrustes distance (4.10) is found between the estimated configurations, then average linkage cluster analysis uses these distances to group the configurations together. The labelling denotes which transform function has been used $E$ denotes the exponential transform (4.4) and $P$ denotes the power transform (4.6); then $M$ denotes metric MDS and $NM$ denotes non-metric MDS. For example $(P, M)$ is the estimated chromosome configuration found using the power transform and metric MDS; $(E, NM)$ is the estimated chromosome configuration found using the exponential transform and non-metric MDS.

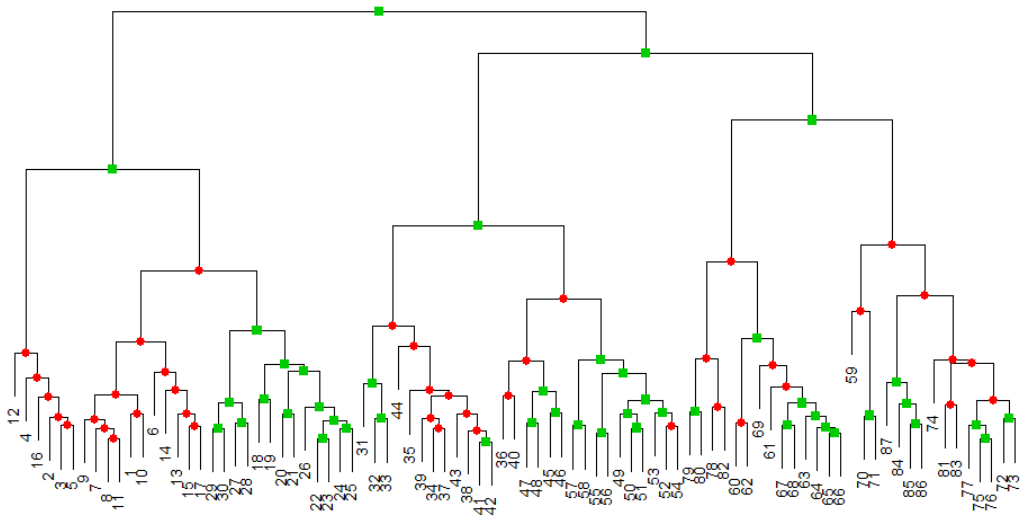
with an indicator system, to indicate if the points clustering together could be ordered sequentially without any missing elements. The indicator system worked as follows: if the total elements of two branches aggregating at the node could be ordered sequentially without any missing elements then a green square ▪ was place on the node, otherwise a red circle was placed on the node ●. For example if the total elements of two branches aggregating at the node are 1,2,3 and 5 then ● is placed on the node; if the total elements of two branches aggregating at the node are 1,2,3,4 and 5 then ▪ is placed on the node.

The dendogram from $\hat{\mathbf{D}}_{E,M}$ in Figure 4.15a contains three large clusters each containing sequential elements, each can be subdivided into two non-sequential subclusters, clusters one contains megabase intervals $1-30$, cluster two contains megabase intervals $31-58$ and cluster contains three megabase intervals $59-87$. The dendogram from $\hat{\mathbf{D}}_{P,M}$ in Figure 4.15b contains two large non-sequential clusters and one small non-sequential cluster.

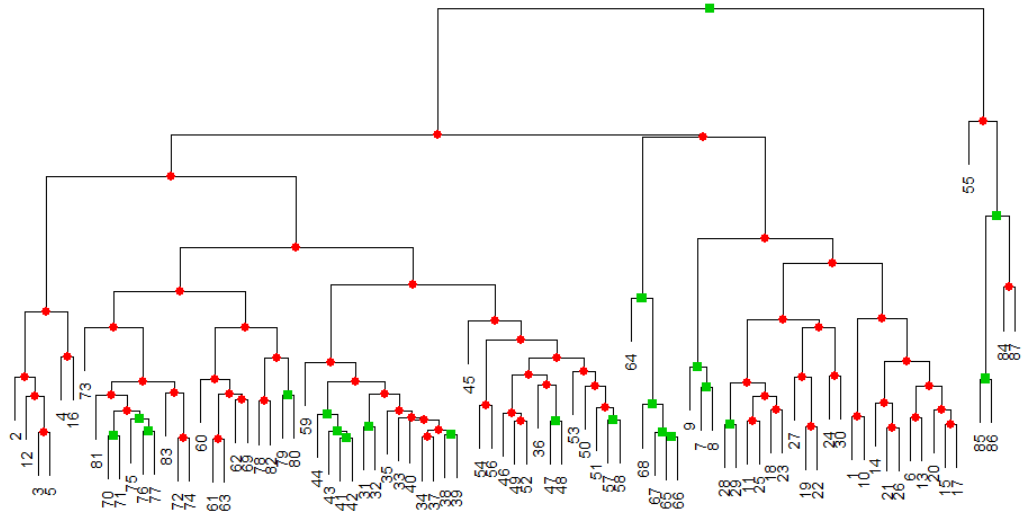
### 4.4.2 Discussion

The elements of the three large clusters on $\hat{\mathbf{D}}_{E,M}$ (Figure 4.15a) correspond to the three blocks on the diagonal of $\hat{\mathbf{D}}_{E,M}$, marking the boundaries of the clusters in $\mathbf{M}$ in Figure 4.16b. They partition the matrix exactly where the blocks abut. Similar partitioning is not possible for the clusters of $\hat{\mathbf{D}}_{P,M}$ (Figure 4.15b) as the clusters are non-sequential. For the same reason of non-sequentiality, the subclusters of $\hat{\mathbf{D}}_{E,M}$ cannot partition $\mathbf{M}$.

The clusters labelled on $\hat{\mathbf{X}}_{E,M}$ in Figure 4.16a could correspond to a higher level of chromatin organisation after the chromatin globule (Sanyal et al., 2011), where globules are brought together to form chromatin domains. Since the elements of the clusters are sequential this gives $\hat{\mathbf{X}}_{E,M}$ a beads on a string appearance, with the clusters forming the beads and the megabase intervals on the boundaries of the clusters the string. Megabase interval $59$ is a example of a boundary megabase interval. It is the last element to be



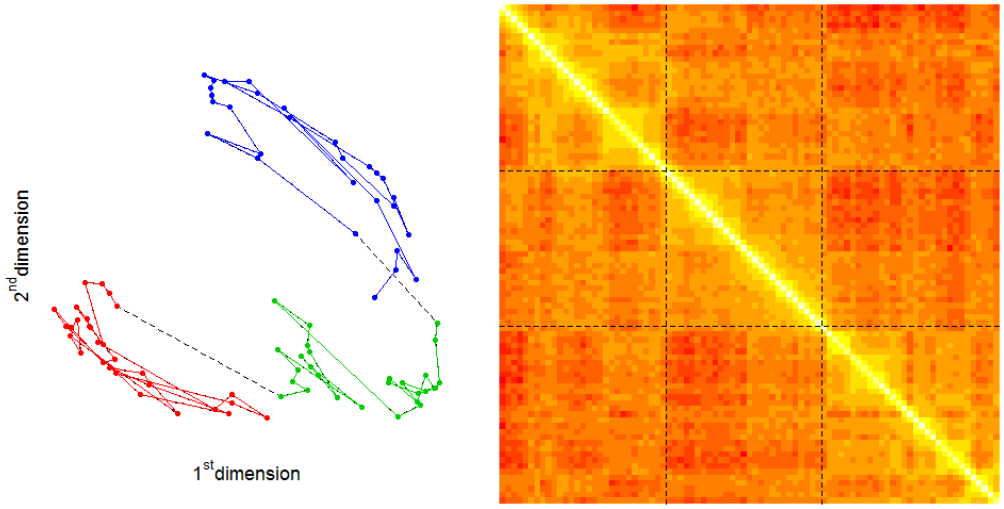
(a) Dendogram from $\hat{\mathbf{D}}_{E,M}$ found using average linkage clustering.



(b) Dendogram from $\hat{\mathbf{D}}_{P,M}$ found using average linkage clustering.

Figure 4.15: Top dendogram: found from the fitted distance matrix $\hat{\mathbf{D}}_{E,M}$ using average linkage clustering. Bottom dendogram: found from the fitted distance matrix $\hat{\mathbf{D}}_{P,M}$ using average linkage clustering. The square ■ indicates if the total elements of the two branches aggregating at a node can be ordered sequentially without any missing elements; otherwise the circle ● is used. The matrix $\hat{\mathbf{D}}_{E,M}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$ using (2.2) with $k = 3$, where $\hat{\mathbf{X}}_{E,M}$ is found in Section 4.2 using the exponential transform (4.4) and metric MDS. The matrix $\hat{\mathbf{D}}_{P,M}$ is extracted from Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$ using (2.2) with $k = 3$, where $\hat{\mathbf{X}}_{P,M}$ is found in Section 4.2 using the power transform (4.6) and metric MDS.

added to cluster three, and the first megabase interval within the cluster. This indicates megabase interval 59 could act as a bridge between clusters two and three. This beads on a string configuration agrees with the fractal globule model (Section 1.2.6) proposed by Lieberman-Aiden et al. (2009) where chromatin clusters can uncrumple without been affected by other chromatin clusters. Although the subclusters do not conform with this fractal globule model as they are non-sequential, these could be driven by the euchromatin and heterochromatin compartments identified by Lieberman-Aiden et al. (2009). Therefore Metric MDS on $\tilde{\mathbf{D}}_{E,M}$ (and non-metric MDS on $\tilde{\mathbf{D}}_{E,NM}$ or $\tilde{\mathbf{D}}_{P,M}$) has potentially captured some local structure from $\mathbf{X}$ in $\hat{\mathbf{X}}_{E,M}$.




(a) Clusters in $\hat{\mathbf{X}}_{E,M}$. (b) Partitioned $\mathbf{M}$.

Figure 4.16: Left panel: the three large clusters identified in Figure 4.15a marked on to Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$. Right panel: boundaries between the three clusters identified in Figure 4.15a partitioning Chromosome 14's Hi-C count matrix. In Figure 4.16a megabase intervals of the first cluster are denoted by ⟶; the megabase intervals of the second cluster are denoted by ⟶, and the megabase intervals of the third cluster are denoted by ⟶. The dashed line - - - denotes the intervals between the clusters. The estimated configuration $\hat{\mathbf{X}}_{E,M}$ is found using the expontial transform (4.4) and metric MDS on Chromosome 14's Hi-C count matrix $\mathbf{M}$ (Section 4.2.1).

## 4.5 Local structure

Small counts $m_{i,j}$ found at large genomic distance provide less accurate estimated distances $\tilde{d}_{i,j}$, which can be detrimental in the recovery $\hat{\mathbf{X}}$. Taking symmetric submatrices from the diagonal of $\mathbf{M}$, contain large $m_{i,j}$ made at low genomic distance and fewer small $m_{i,j}$. These submatrices can be transformed and fitted using the same approaches used to fit $\mathbf{M}$ to give an estimated configuration for a subregion of the chromosome. Two approaches to obtain estimated configurations from submatrices were taken.

Let $\mathbf{M}^{(a:b)} = (m_{i,j}^{(a:b)})$ denote the symmetric submatrix of $\mathbf{M}$ from megabase interval $a$ through to megabase interval $b$, such that $m_{i,j}^{(a:b)} = m_{i+a-1,j+b-1}$, the same superscript notation is added to all matrices involved.

### 4.5.1 Single submatrices

This method takes a single symmetric submatrix $\mathbf{M}^{(a:b)}$, transforms it into estimated distance matrix $\tilde{\mathbf{D}}^{(a:b)}$ using (4.4) or (4.6). Then $\tilde{\mathbf{D}}^{(a:b)}$ is fitted into Euclidean space using metric or non-metric MDS. The parameters $\hat{\alpha}$ for (4.4) or $\hat{\beta}$ for (4.6) are found using the fitting algorithm in Section 4.1.3. Choosing the correct size $\mathbf{M}^{(a:b)}$ is important. Too small a $\mathbf{M}^{(a:b)}$ then little new information will be gained. For example a $2 \times 2$ $\mathbf{M}^{(a:b)}$ will provide no new information. Too large a $\mathbf{M}^{(a:b)}$ then again little new information will be gained. For example a $(n-1) \times (n-1)$ $\mathbf{M}^{(a:b)}$ will give little new information. Medium sized $\mathbf{M}^{(a:b)}$ could provide some new information. Medium $\mathbf{M}^{(a:b)}$ will not be as influenced by small counts found at large genomic distance, and instead will contain a higher proportion of the larger counts found at small genomic distance.

The elements of the three sequential clusters identified from $\hat{\mathbf{D}}_{E,M}$ for Chromosome 14 in Section 4.4, provides the megabase intervals to extract $\mathbf{M}^{(a:b)}$. These three $\mathbf{M}^{(a:b)}$ will be transformed into estimated distances $\tilde{\mathbf{D}}_{E,M}^{(a:b)}$ using the exponential transform (4.4),

and fitted into three dimensional Euclidean space using metric MDS, to give the fitted configurations $\hat{\mathbf{X}}_{E,M}^{(a:b)}$. The $\hat{\alpha}$ parameters which minimize the score function $\chi^2$ (4.9) will be found using the fitting algorithm in Section 4.1.3.

| Cluster | Megabase intervals | $\hat{\alpha}$ estimate | $\chi^2$ |
|---|---|---|---|
| 1 | 1:30 | 0.0058 | 12370 |
| 2 | 31:58 | 0.0046 | 16576 |
| 3 | 59:87 | 0.0047 | 22120 |

Table 4.7: Score function data from fitting single submatrices $\mathbf{M}^{(a:b)}$ from Chromosome 14's Hi-C counts matrix, using the exponential transform (4.4) and metric MDS. Column one and two identify which large cluster (Section 4.4) the submatrix $\mathbf{M}^{(a:b)}$ corresponds to, and the megabase intervals $a : b$ it is composed of. Column three and four give the estimated parameter $\hat{\alpha}$ value and the $\chi^2$ (4.9) value it minimizes. The $\hat{\alpha}$ for the exponential transform is found by applying the fitting algorithm (in Section 4.1.3) using the $\chi^2$ score function and fitting into three dimensional Euclidean space with metric MDS, to $\mathbf{M}^{(a:b)}$.

The fitted configurations $\hat{\mathbf{X}}_{E,M}^{(1:30)}, \hat{\mathbf{X}}_{E,M}^{(31:58)}$ and $\hat{\mathbf{X}}_{E,M}^{(59:87)}$ from the three submatrices all resemble smaller copies of $\hat{\mathbf{X}}_{E,M}$ (Figure 4.5). The points in the first and second dimensions when plotted resemble a horseshoe shape, and the points in the first and third dimensions resemble a cubic polynomial. This indicates that the horseshoe effect (Section 2.4) is present in $\hat{\mathbf{X}}_{E,M}^{(1:30)}, \hat{\mathbf{X}}_{E,M}^{(31:58)}$ and $\hat{\mathbf{X}}_{E,M}^{(59:87)}$. The centres of the fitted configurations also appear hollow. The fitted configuration of the second cluster $\hat{\mathbf{X}}_{E,M}^{(31:58)}$ can be seen in Figure 4.17.

### 4.5.2 Windows smoothing

Windows smoothing is an advancement on the single submatrices method. Windows smoothing takes all the $n^* \times n^*$ (where $n^* < n$) symmetric submatrices from $\mathbf{M}$, stacks the submatrices into an $n^* \times n^* \times (n - n^* + 1)$ array, finally the mean is taken down the

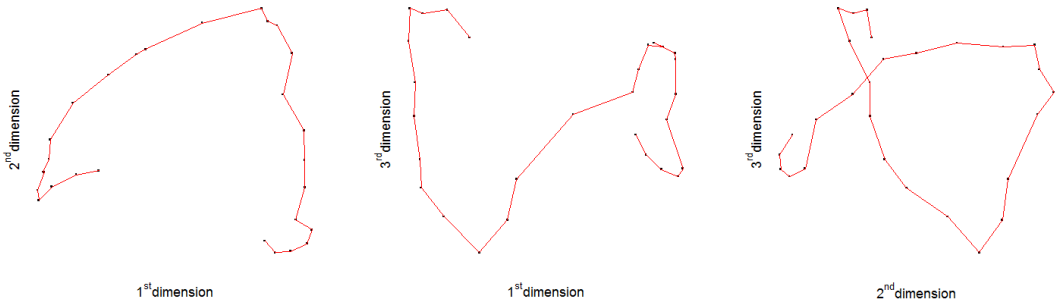



Figure 4.17: Perspectives of Chromosome 14's estimated configuration for the second cluster $\hat{\mathbf{X}}_{E,M}^{(31:58)}$. The origin of the megabase interval is denoted by . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}_{E,M}^{(31:58)}$ is found by taking the submatrix $\mathbf{M}^{(31:58)}$, transforming it into estimated distances using the exponential transform (4.4) with $\hat{\alpha} = 0.0046$ (Table 4.7). Then fitting the estimated distances into three dimensional Euclidean space with metric MDS. The single submatrix $\mathbf{M}^{(31:58)}$ is taken of megabase intervals 31 to 58 from Chromosome 14's Hi-C count matrix.

plates of the array to give a windows smoothed submatrix $\mathbf{W}^{(n^*)} = (w_{i,j}^{(n^*)})$ where

$$w_{i,j}^{(n^*)} = (n - n^* + 1)^{-1} \sum_{k=0}^{n-n^*} m_{i+k,j+k} \text{ for } i = 1, \ldots, n^*,\ j = 1, \ldots, n^* \tag{4.12}$$

Windows smoothing intends to capture mean local chromosome structure, by minimizing influence from the noise in small $m_{i,j}$.

Taking a window sized $15 \times 15$ megabase intervals from Chromosome 14's Hi-C count matrix $\mathbf{M}$, the windows smoothed submatrix $\mathbf{W}^{(15)}$ is transformed and fitted using each of the two transform functions (4.4) and (4.6) and metric MDS. The score function $\chi^2$ (4.9) is minimized using $\mathbf{W}^{(15)}$ and the fitted windows smoothed counts $\hat{\mathbf{W}}^{(15)}$. The $m_{\text{min}} = 1$ adjustment for the power transform is unnecessary as $w_{i,j}^{(15)} > 2\ \forall i \neq j$ (all counts in $\mathbf{W}^{(15)}$ are larger than 2). Table 4.8 shows that the power transform (4.6) has produced a better fitting configuration with a smaller $\chi^2$, than the corresponding value for the exponential transform (4.4). The fitted configurations $\hat{\mathbf{X}}_{E,M}^{(15)}$ and $\hat{\mathbf{X}}_{P,M}^{(15)}$ from

| Transform function | Parameter | Parameter estimate | $\chi^2$ |
|---|---|---|---|
| Exponential transform | $\hat{\alpha}$ | 0.0035 | 1244 |
| Power transform | $\hat{\beta}$ | -0.7451 | 687 |

Table 4.8: Score function data from fitting a windows smoothed submatrix sized $15 \times 15$ megabases $\mathbf{W}^{(15)}$ from Chromosome 14's Hi-C counts matrix. Column one and two state which transform function has been used and which parameter has been estimated. Column three and four give the estimated parameter value and the $\chi^2$ (4.9) value it minimizes. The $\hat{\alpha}$ for (4.4) and $\hat{\beta}$ for (4.6) are found by applying the fitting algorithm (in Section 4.1.3) using the $\chi^2$ (4.9) score function and fitting into three dimensional Euclidean space with metric MDS, to $\mathbf{W}^{(15)}$.

$W^{(15)}$ displayed in Figures 4.18 and 4.19 take a horseshoe shape in the first and second dimensions, and a cubic polynomial shape in the first and third dimensions. The fitted configuration for the power transform $\hat{\mathbf{X}}^{(15)}_{P,M}$ appears to have taken on less aspects the horseshoe effect than $\hat{\mathbf{X}}^{(15)}_{E,M}$. It has minor distortion in the points and no involution at the ends of the configuration.

### 4.5.3 Discussion

All the fitted configurations (found from submatrices and windows smoothing) obtained to investigate local structure, appear to have taken on aspects of the horseshoe effect. Using smaller matrices has avoided the smaller counts found at large genomic distance, which produce large spatial distances. This has caused the medium spatial distances to become the new large distances of the estimated distance matrix, allowing the horseshoe effect to be repeated. This provides evidence that the transform functions or score functions used to obtain $\hat{\mathbf{X}}$, $\hat{\mathbf{X}}^{(a:b)}$ or $\hat{\mathbf{X}}^{(n^*)}$, require modification to avoid the presence of the horseshoe effect in the fitted configurations.

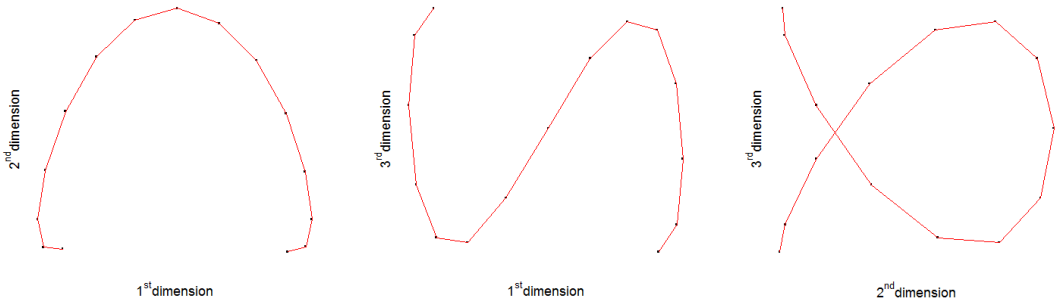



Figure 4.18: Perspectives of Chromosome 14's estimated windows smoothed configuration $\hat{\mathbf{X}}^{(15)}_{E,M}$. The origin of the megabase interval is denoted by . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}^{(15)}_{E,M}$ is found by taking the windows smoothed submatrix $\mathbf{W}^{(15)}$, transforming it into estimated distances using the exponential transform (4.4) with $\hat{\alpha} = 0.0035$ (Table 4.8). Then fitting the estimated distances into three dimensional Euclidean space with metric MDS. The windows smoothed submatrix $\mathbf{W}^{(15)}$ is taken from Chromosome 14's Hi-C count matrix using (4.12) with a window sized $15 \times 15$ megabases.

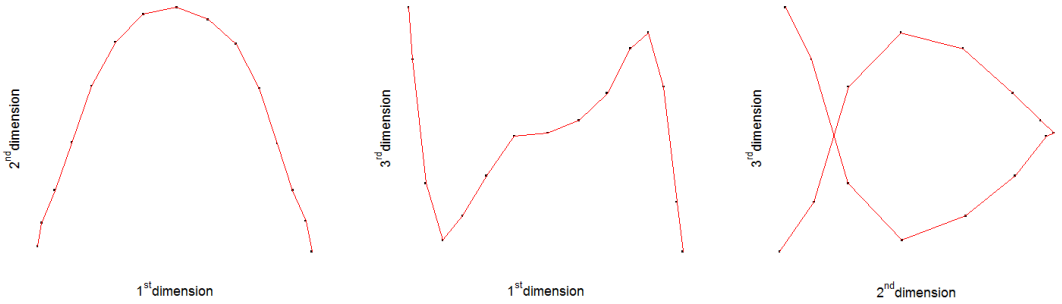



Figure 4.19: Perspectives of Chromosome 14's estimated windows smoothed configuration $\hat{\mathbf{X}}^{(15)}_{P,M}$. The origin of the megabase interval is denoted by . and the red line — denotes the average path of the DNA along the configuration. The configuration $\hat{\mathbf{X}}^{(15)}_{P,M}$ is found by taking the windows smoothed submatrix $\mathbf{W}^{(15)}$, transforming it into estimated distances using the power transform (4.6) with $\hat{\beta} = -0.7451$ (Table 4.8). Then fitting the estimated distances into three dimensional Euclidean space with metric MDS. The windows smoothed submatrix $\mathbf{W}^{(15)}$ is taken from Chromosome 14's Hi-C count matrix using (4.12) with a window sized $15 \times 15$ megabases.

## 4.6 Additional properties influencing count size

The main assumption underlying the recovery of a true chromosome configuration $\mathbf{X} = (x_{i,k})$ is $m_{i,j}$ are related to $d_{i,j}$, through some unknown relationship and with the addition of some noise, although other properties of the chromatin influence the $m_{i,j}$. The influence of euchromatin and heterochromatin is one property.

Genomic distance $\mathbf{N} = (\nu_{i,j})$ measures the distance in megabases (Mb) between megabase intervals along the chromosome. Adjacent megabase intervals $i$ and $i+1$ have a genomic distance $\nu_{i,i+1} = 1$. Non-adjacent megabase intervals $i$ and $j$ have a genomic distance $\nu_{i,j} = |i - j|$. Genomically close megabase intervals $i$ and $j$ are also spatially close and tend to have large $m_{i,j}$. Imagine each megabase interval as a link in a chain: even when the chain is taut, close links along the chain remain spatially close.

The degree to where chromatin is condensed can also affect the size of $m_{i,j}$. Euchromatin is less condensed and is free to make more contacts, whereas heterochromatin is more condensed and makes fewer contacts.

### 4.6.1 Generalized linear regression

Generalized linear regression with a logarithmic link function was used to investigate the direct relationship of $m_{i,j}$ with $\nu_{i,j}$ (Christensen (1990) pages 349-364, Dobson and Barnett (2008) pages 165 - 183). Two models were used. The first model (4.13) (Model A) using $\log(\nu_{i,j})$ as the explanatory variable, the second model (4.14) (Model B) augmenting the first model (4.13) to include row $\underline{\phi} = (\phi_i)$ and column $\underline{\psi} = (\psi_j)$ factors which could be used to interpret the megabase effect. The megabase effect is how the proportions of euchromatin or heterochromatin which make up the interval influence $m_{i,j}$. If a megabase intervals main constituent is euchromatin, then it should be more open and make contact frequently, resulting in larger $m_{i,j}$. If a megabase intervals main

constituent is heterochromatin, then it should be more close and make fewer contacts, resulting in smaller $m_{i,j}$.

**Model A**

$$\log(E(m_{i,j})) = c_0 + c_1 \log(\nu_{i,j}). \tag{4.13}$$

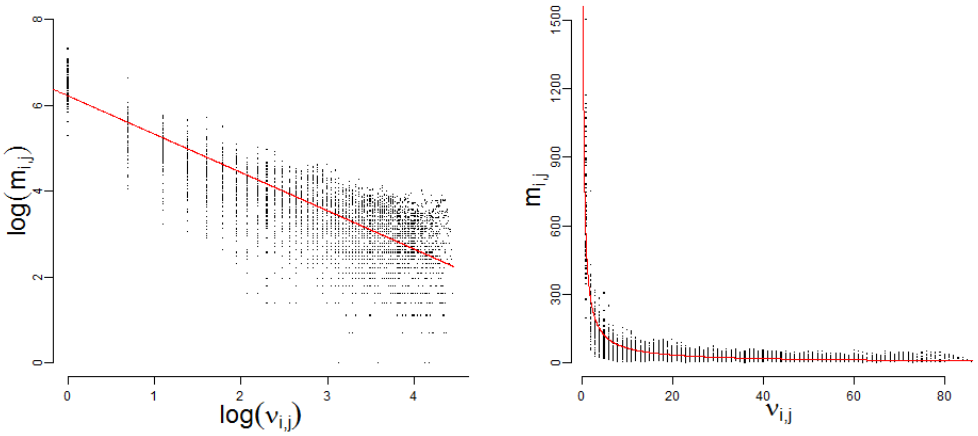
Model A produces a monotonically decreasing relationship between $m_{i,j}$ and $\nu_{i,j}$ (since

| | Estimate | Standard error |
|---|---|---|
| $c_0$ | 6.2222 | 0.0153 |
| $c_1$ | -0.8923 | 0.007 |
| $\hat{\rho}$ | 17.2594 | |

Table 4.9: Coefficient estimates with their standard errors, and the dispersion estimate $\hat{\rho}$; from applying Model A (4.13) to Chromosome 14's Hi-C count matrix.

$c_1 < 0$). Figure 4.20 shows a steep decline in $m_{i,j}$ at $1$ Mb $\leq \nu_{i,j} < 15$ Mb then a steadier decline in $m_{i,j}$ at $\nu_{i,j} \geq 15$ Mb, this is in agreement with Dekker et al. (2013) description of the count and genomic distance relationship. Dekker et al. (2013) describes Hi-C (Lieberman-Aiden et al., 2009) data having a steep decline for contact probability at genomic distance $1$ Mb $\leq \nu_{i,j} < 10$ Mb, and a shallow decline for contact probability at genomic distance $\nu_{i,j} > 10$ Mb. The presence of an underlying gradient gives M an almost Toeplitz structure (see Section 2.4) and could be the cause of the horseshoe shapes present in $\hat{\mathbf{X}}$ Figures 4.5, 4.6, 4.12 and 4.13. De Leeuw (2008) discusses how Toeplitz dissimilarity matrices produce horseshoes when fitted using metric MDS. The transition from steep to steady gradients Figure 4.20b indicates where the blending of medium and large distance begins and where it becomes difficult to accurately measure distance. For example, megabase intervals 1 and 21 share a similar $E(m_{i,j})$ to megabase intervals 1 and 71, $E(m_{1,21}) \approx E(m_{1,71})$, due to the plateauing of $E(m_{i,j})$ at $\nu_{i,j} \geq 15$ Mb.

The underlying gradient and plateau at $\nu_{i,j} \geq 15$Mb are intrinsic factors in $\mathbf{M}$ and complicate finding a suitable transform and score function.




(a) $\log(m_{i,j})$ against $\log(\nu_{i,j})$. (b) $m_{i,j}$ against $\nu_{i,j}$.

Figure 4.20: Plots of Chromosome 14's Hi-C counts $\mathbf{M} = (m_{i,j})$ against genomic distance $\mathbf{N} = (\nu_{i,j})$. Left panel: $\log(m_{i,j})$ against $\log(\nu_{i,j})$. Right panel: $m_{i,j}$ against $\nu_{i,j}$. In both plots the red line — denotes the line of best fit from Model A (4.13).

**Model B**

$$\log(E(m_{i,j})) = c_1 \log(\nu_{i,j}) + \phi_i + \psi_j, \tag{4.14}$$

where $\underline{\phi} = (\phi_i)$ and $\underline{\psi} = (\psi_i)$ are the row and column effects and $i, j = 1, \ldots, n$. Since

| | Estimate | Standard error |
|---|---|---|
| $c_1$ | -1.1660 | 0.0104 |
| $\hat{\rho}$ | 10.2538 | |

Table 4.10: Coefficient estimates with their standard errors, and the dispersion estimate $\hat{\rho}$; from applying Model B (4.14) to Chromosome 14's Hi-C count matrix.

$\mathbf{M}$ is a symmetric matrix, the row and column effects are two incomplete parts of the same

piece of information. Combining $\underline{\phi}$ and $\underline{\psi}$ will provide the complete piece of information, a vector of megabase effects $\underline{\pi} = (\pi_i)$, where $\pi_i = \phi_i + \psi_i$. The diagonal entries of $\mathbf{M}$ (intramegabase counts) also provide information on the level of condensing in a megabase interval, larger $m_{i,i}$ would indicate the intervals main constituent is euchromatin and smaller $m_{i,i}$ would indicate the intervals main constituent is heterochromatin. Given this the diagonal entries of $\mathbf{M}$, $\underline{\omega} = (\omega_i)$ where $\omega_i = m_{i,i}$, are compared with $\underline{\pi}$ in Figure 4.21. The comparison is looking to see if the distribution of peaks and troughs in $\underline{\pi}$ is similar to $\underline{\omega}$.

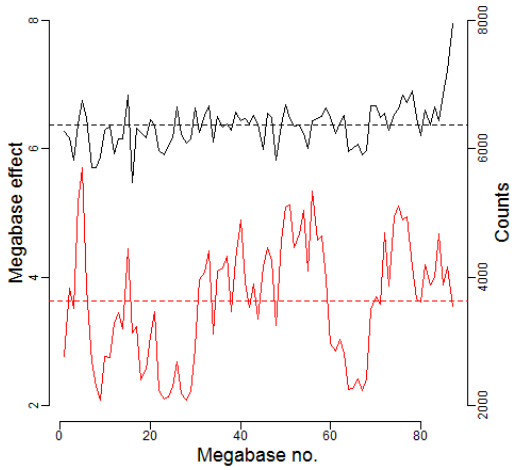



Figure 4.21: Comparison of the megabase effect $\underline{pi} = (\pi_i)$ with the intra-megabase counts $\underline{\omega} = (\omega_i)$, for Chromosome 14's Hi-C count matrix $\mathbf{M} = (m_{i,j})$. The megabase effects $\underline{\pi} = (\pi_i)$ are denoted by the black line ——, with the left axis giving its values. The intra megabase counts $\underline{\omega} = (\omega_i)$ are denoted by the red line ——, with the right axis giving its values. The $\underline{\pi}$ values are found by summing the row and column effects (4.14), $\underline{\pi} = \underline{\phi} + \underline{\psi}$. The $\underline{\omega}$ values are the diagonal entries of $\mathbf{M}$, $\underline{\omega} = \text{diag}(\mathbf{M})$.

The distribution of peaks and troughs in Figure 4.21 looks broadly similar for $\underline{\pi}$ and $\underline{\omega}$, suggesting either measure could be used to assess the level of condensing within a megabase. The $\underline{\omega}$ values were used to determine if a megabase intervals tended produce large $m_{i,j}$'s and was more euchromatin-like (extrovert) or small $m_{i,j}$'s and was more heterochromatin-like (introvert). If $\omega_i > \overline{\omega}$ where $\overline{\omega} = n^{-1} \sum_{i=1}^{n} \omega_i$ then megabase $i$ was

relatively extrovert otherwise if $\omega_i \leq \overline{\omega}$ then megabase $i$ was relatively introvert.

In Figures 4.22 and 4.23 the estimated chromosome configurations $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,M}$ for Chromosome 14, have had the extrovert and introvert megabase intervals labelled on them. In both estimated chromosome configurations the introverted megabase intervals cluster together and the extroverted megabase intervals cluster together. Although the two clusters appear to be almost separated, by some imaginary partition.

The positioning of extroverted and introverted megabases in $\hat{\mathbf{X}}_{E,M}$ (Figure 4.22) and $\hat{\mathbf{X}}_{P,M}$ (Figure 4.23) could correspond to the two compartments observed by Lieberman-Aiden et al. (2009). The extroverted and introverted megabase intervals in $\hat{\mathbf{X}}_{P,M}$ broadly match with the clusters identified in the dendogram of $\hat{\mathbf{D}}_{P,M}$ in Figure 4.15b. Clusters one and three broadly match the extroverted megabase intervals and cluster two broadly matches the introverted megabase intervals. This matching indicates that the cluster analysis in $\hat{\mathbf{D}}_{P,M}$ detects the compartment feature of the genome.

### 4.6.2 Discussion

Investigating the relationship between $m_{i,j}$ and $\nu_{i,j}$, has drawn attention to the gradient in how count size declines with increasing genomic distance in $\mathbf{M}$ and how the MDS fits megabase intervals dependent on their level of condensing. The gradient in $\mathbf{M}$ implies that preprocessing or appropriate transform and score functions must be developed to remove the horseshoe effect in $\hat{\mathbf{X}}$.

## 4.7 Normalization and filtering

This section looks at ideas to normalize $\mathbf{M}$ and filter out small counts from $\mathbf{M}$. The power transform (4.6) requires all observed counts of zero to be adjusted to one, to provide a distance matrix which can be fitted into Euclidean space. Alternative methods

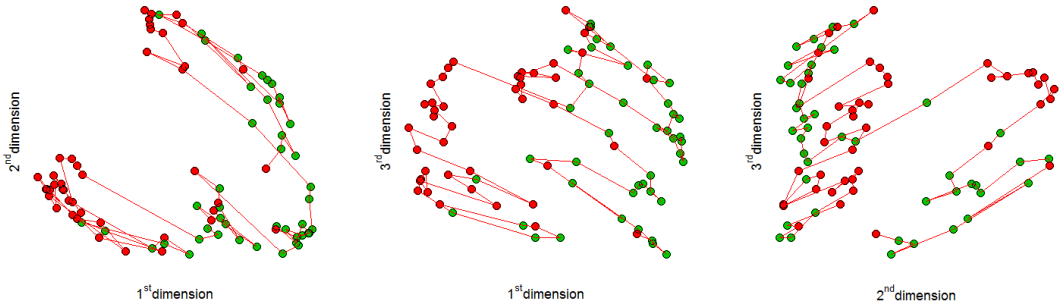



Figure 4.22: Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{E,M}$ found using the exponential transform (4.4) and metric MDS, with extroverted and introverted megabase intervals labelled. The green circle ● denotes if the megabase interval is extroverted and the red circle ● denoted if the megabase interval is introverted (Section 4.6). The configuration $\hat{\mathbf{X}}_{E,M}$ is found in Section 4.2.1 from Chromosome 14's Hi-C count matrix.

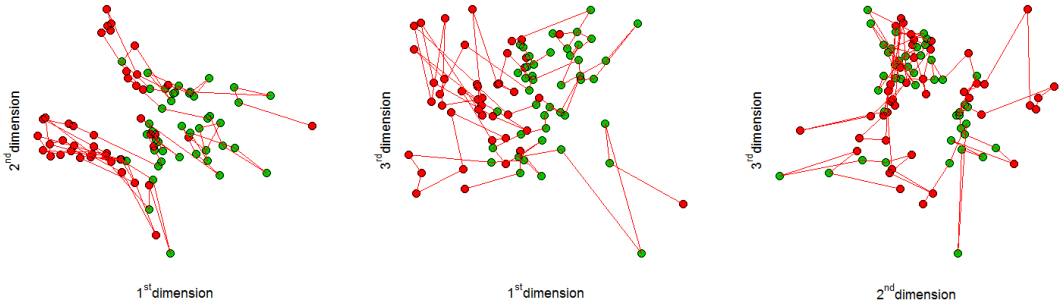



Figure 4.23: Chromosome 14's estimated configuration $\hat{\mathbf{X}}_{P,M}$ found using the power transform (4.6) and metric MDS, with extroverted and introverted megabase intervals labelled. The green circle ● denotes if the megabase interval is extroverted and the red circle ● denoted if the megabase interval is introverted (Section 4.6). The configuration $\hat{\mathbf{X}}_{P,M}$ is found in Section 4.2.1 from Chromosome 14's Hi-C count matrix.

of removing small counts from $\mathbf{M}$ are explored as an alternative to the $m_{\text{min}}$ adjustment. Normalization and filtering aim to preprocess $\mathbf{M}$, fit the preprocessed $\mathbf{M}$ with the power transform and metric MDS, to recover $\hat{\mathbf{X}}_{P,M}$ with a less chaotic pattern and remove the need for the $m_{\text{min}}$ adjustment. The normalization procedure used by Lieberman-Aiden et al. (2009); Duan et al. (2010); Trieu and Cheng (2014) produces a matrix of "interaction frequencies" $\mathbf{F} = (f_{i,j})$ giving the ratio of observed to expected count,

$$f_{i,j} = m_{i,j} \times \left( \frac{\sum_{i=1}^{n} m_{i,j} \sum_{j=1}^{n} m_{i,j}}{\sum_{i<j} m_{i,j}} \right)^{-1}, \tag{4.15}$$

where the second term (inside the brackets) is the reciprocal of the expected count. Yaffe and Tanay (2011) normalize $\mathbf{M}$ for experimental biases in the data caused by distances between restriction sites in the chromosome; guanine and cytosine (GC) content of trimmed ligation junctions and sequence uniqueness. Where the distance between restriction sites, is how the lengths of the fragments produced in the Hi-C experiments are related to intrachromosomal and interchromosomal contacts. For example short fragments might produce more intrachromosomal contacts and fewer interchromosomal contacts. The GC content bias is a bias driven by the nucleotide composition of the studied DNA. The GC content near restriction ends has an effect on the probability of interchromosomal contacts been made. Imakaev et al. (2012) applies an iterative decomposition to $\mathbf{M}$ to correct for biases and potential future biases.

### 4.7.1 Observed count normalization

An alternative normalization uses the diagonal elements of $\mathbf{M}$ to normalize $\mathbf{M}$, to give a normalized observed count matrix $\mathbf{M}^* = (m^*_{i,j})$ with the intention of reducing the extrovert-introvert megabase effect observed in Section 4.6. To begin with we constructed

a diagonal matrix $\mathbf{Q} = (q_{i,i})$, such that

$$q_{i,i} = \sqrt{\frac{n^{-1}\sum_{i=1}^{n} m_{i,i}}{m_{i,i}}}.$$

Then obtained the normalized counts $\mathbf{M}^*$ though

$$\mathbf{M}^* = \mathbf{QMQ}. \tag{4.16}$$

The normalized count matrix $\mathbf{M}^*$ was transformed and fitted using the power transform (4.6) and metric MDS, although retaining the $m_{\text{min}}$ adjustment.

The $\chi^2$ (4.9) results from the normalization in Table 4.11 display an improvement for the $m_{\text{min}} = 1$ adjustment and a deterioration for the $m_{\text{min}} = 2$ adjustment. The fitted configuration $\hat{\mathbf{X}}^*_{P,M}$ for $m_{\text{min}} = 2$ appears similar to $\hat{\mathbf{X}}_{P,M}$ in Figure 4.6, so normalization has failed to reduce the extrovert introvert megabase effect.

| Adjustment | Normalized $\chi^2$ |
|---|---|
| $m_{\text{min}} = 1$ | 1060582 |
| $m_{\text{min}} = 2$ | 533662 |

Table 4.11: Score function data from fitting Chromosome 14's normalized Hi-C count matrix $\mathbf{M}^*$ (4.16), with the power transform (4.6) and metric MDS. Column one states which count adjustment has been applied to $\mathbf{M}^*$. Column two gives the minimized $\chi^2$ (4.9) values. The minimized $\chi^2$ values are found by applying the fitting algorithm (Section 4.1.3) with the power transform and fitting into three dimensional Euclidean space with metric MDS, to $\mathbf{M}^*$.

### 4.7.2 Observed count filtering

Filtering uses a generalized linear regression model with logarithmic link function, to filter small $m_{i,j}$ from $\mathbf{M}$, as an alternative to the $m_{\text{min}}$ adjustment. Using model A (4.13)

a matrix of expected counts $E(\mathbf{M}) = (E(m_{i,j}))$ is generated with an estimate for the dispersion $\hat{\rho} = 17.2594$ in $\mathbf{M}$ obtained from the model, where var$(m_{i,j}) = \rho\mu_{i,j}$. Using the negative binomial distribution (suited for over dispersed counts) with expected counts $E(m_{i,j})$ and dispersion $\hat{\rho}$, a matrix of probabilities $\mathbf{P} = (p_{i,j})$ is constructed such that $p_{i,j} = P(m_{i,j} \leq E(m_{i,j})|E(m_{i,j}), \hat{\rho})$. The $p_{i,j}$ gives the probability of observing a count $m_{i,j}$ less than or equal to the expected count $E(m_{i,j})$. A threshold probability is chosen $\tau$ and a corresponding matrix of mean counts $\mathbf{T} = (t_{i,j})$ at this probability are found using the negative binomial distribution quantile function, such that $t_{i,j}$ is the lowest integer that allows $P(t_{i,j} \leq E(m_{i,j})|E(m_{i,j}), \hat{\rho}) \geq \tau$. Then any $m_{i,j}$ with a probability $p_{i,j} < \tau$, are replaced with the corresponding $t_{i,j}$, to give a filtered count matrix $\tilde{\mathbf{M}} = (\tilde{m}_{i,j})$. The idea behind this is if $m_{i,j}$ is classified as been too small according to Model A (4.13), then it can be replaced with a value that is just large enough according to the model. This avoids drastically increasing the size of the counts. The filtered counts $\tilde{\mathbf{M}}$ are then transformed and fitted using the power transform (4.6) and metric MDS. Three threshold values were used: $\tau = 0.05,\ 0.01$ and $0.005$. In each case the $m_{\text{min}}$ adjustment was retained, so filtering failed.

Only at $\tau = 0.05$ did filtering improve $\chi^2$ (4.9), while $\tau = 0.01$ and $0.005$ lead to deterioration in $\chi^2$. Filtering at $\tau = 0.05$ led to the replacement of 191 individual $m_{i,j}$ (382 in total). The locations of the replaced $m_{i,j}$ are plotted in Figure 4.24. All the replacements take place close to the diagonal and no small $m_{i,j}$ were replaced. Therefore filtering has worked counter to its intention by removing large $m_{i,j}$ and leaving small the $m_{i,j}$, with the $m_{\text{min}} = 1$ adjustment still required.

### 4.7.3 Discussion

Preprocessing $\mathbf{M}$ through normalization or filtering improved some $\chi^2$ values, but the improvements in $\hat{\mathbf{X}}_{P,M}$ were not noticeable or the intention of the preprocessing was not fulfilled. Future normalization of $\mathbf{M}$ could use the interaction frequency matrix (4.15)

| Threshold level | $\chi^2$ |
|---|---|
| $\tau = 0.05$ | 858164 |
| $\tau = 0.01$ | 1746783 |
| $\tau = 0.005$ | 2614803 |

Table 4.12: Score function data from fitting Chromosome 14's filtered Hi-C count matrix $\tilde{\mathbf{M}}$, with the power transform (4.6) and metric MDS. Column one gives the threshold level $\tau$ used in the filtering (Section 4.7.2) to obtain $\tilde{\mathbf{M}}$. Column two gives the minimized $\chi^2$ (4.9) values. The minimized $\chi^2$ (4.9) values are found by applying the fitting algorithm (Section 4.1.3) with the power transform (4.6) and fitting into three dimensional Euclidean space with metric MDS, to $\tilde{\mathbf{M}}$ (Section 4.7.2).

used by other research groups while future filtering could fit a generalized linear model truncated to model the counts made at large genomic distance $\nu_{i,j} \leq 15Mb$. This would avoid filtering out the large counts made at low genomic distance, which are not detrimental to the estimated distances or estimated chromosome configuration.

## 4.8 Estimated genome configuration

In addition to obtaining estimated chromosome configurations, the global contact matrix can be transformed and fitted into three dimensional Euclidean space, to give an estimated genome configuration. The global contact matrix $\mathbf{G}^{(r)} = (g_{i,j}^{(r)})$ (where the superscript notation $r$ denotes the resolution in megabases (Mb)), contains all the intrachromosomal and interchromosomal counts for the twenty two chromosome pairs and the XX pair. At 1 Mb resolution, with the rows and columns of poor mapability deleted, $\mathbf{G}^1$ is a $2820 \times 2820$ symmetric matrix. The same method of transforming and fitting $\mathbf{M}$ can be applied to $\mathbf{G}^1$ to obtain a fitted genome configuration $\hat{\mathbf{Z}}^1 = (\hat{z}_{i,k}^{(1)})$ (here $\mathbf{Z}$ is used to denote to the genome configuration to avoid confusion with the chromosome configuration), although at 1 Mb resolution points are susceptible to noise in the large quantities of small interchromosomal counts and are more difficult to interpret. To reduce the influence of

noise and number of points to interpret resolution was lowered, setting $r = 3; 6; 12$ and $24$ Mb. Lowering resolution sums $g_{i,j}^{(1)}$ with low counts and poor information into $g_{i,j}^{(r)}$ with larger counts and richer information. Although lowering the resolution reduces the level of detail available at high resolution. This is an issue if the individual estimated chromosome configurations are to be extracted from the estimated genome configurations for analysis.

### 4.8.1 Lowering resolution

When lowering resolution it is important that the new $g_{i,j}^{(r)}$ do not straddle chromosomes or cross centromeres. Counts $g_{i,j}^{(r)}$ which violate this provide a mixed information from different chromosome arms or different chromosomes. This is particularly an issue if $g_{i,j}^{(r)}$ straddles chromosomes. The process of lowering the resolution from $1$ Mb to $r$ Mb is outlined below.

1. Identify to which chromosome arm each row and column of $\mathbf{G}^{(1)}$ belongs. Trim rows and columns from $\mathbf{G}^{(1)}$ to give $\mathbf{G}^{(1*)} = (g_{i,j}^{(1*)})$ so the length of chromosome arms become divisible by $r$, ensuring the start and ends of the arms are trimmed equally. This step prevents $g_{i,j}^{(r)}$ straddling chromosomes or centromeres.

2. Sum over $r \times r$ submatrices within $\mathbf{G}^{(1*)}$ to produce $\mathbf{G}^{(r)}$

$$g_{p,q}^{(r)} = \sum_{i=(p-1)r+1}^{pr} \sum_{i=(q-1)r+1}^{qr} g_{i,j}^{(1*)}, \tag{4.17}$$

### 4.8.2 Fitted genome configuration

Using the same approach as in Section 4.2 four different estimated genome configurations $\hat{\mathbf{Z}}^{(r)}$ for each resolution $r$ were obtained, with the same subscript notation used to denote the transform function and MDS method ($E =$ exponential transform (4.4); $P =$ power

transform (4.6); $M$ = metric MDS and $NM$ = non-metric MDS). The fitted configuration $\hat{\mathbf{Z}}^{(r)}$ at resolution $r$ Mb, contains the estimated coordinates of all twenty two human chromosome pairs and the XX chromosome pair. The first $n_1^{(r)}$ rows of $\hat{\mathbf{Z}}^{(r)}$ are the coordinates of Chromosome pair 1. Then rows $n_1^{(r)} + \ldots + n_{a-1}^{(r)} + 1$ to $n_a^{(r)}$ of $\hat{\mathbf{Z}}^{(r)}$ are the coordinates of chromosome pair "$a$" for $a = 2, \ldots, 22$ and XX (where $a = 23$ for XX). Below is a summary of how the estimated genome configurations are found:

1. Lower the resolution of the global Hi-C count matrix $\mathbf{G}^{(1)}$ to a desired resolution $r$ Mb, using the process described in Section 4.8.1 to obtain $\mathbf{G}^{(r)}$.

2. Transform $\mathbf{G}^{(r)}$ into an estimated distance matrix $\tilde{\mathbf{D}}_{\cdot,\cdot}^{(r)}$ using either the exponential transform (4.4) or the power transform (4.6). The transform parameters $\hat{\alpha}$ for (4.4) or $\hat{\beta}$ for (4.6) are found using the fitting algorithm (Section 4.1.3), with the score function $\chi^2$ (4.9) if metric MDS is to be used or $S_p(\hat{\mathbf{X}})$ (2.14) if non-metric MDS is to be used.

3. Fit $\tilde{\mathbf{D}}_{\cdot,\cdot}^{(r)}$ into three dimensional Euclidean space using either metric MDS or non-metric MDS, to obtain an estimated genome configuration $\hat{\mathbf{Z}}_{\cdot,\cdot}^{(r)}$.

Take for example $\hat{\mathbf{Z}}_{P,NM}^{(6)}$ is the estimated genome configuration at 6 Mb resolution, which has been found using the power transform and non-metric MDS.

Visual inspection of the estimated genome configurations from metric MDS $\hat{\mathbf{Z}}_{E,M}^{(r)}$ and $\hat{\mathbf{Z}}_{P,M}^{(r)}$ (Figure 4.25), displays two flat clusters one suspended above the other, which is inconsistent with the known chromosomes territories feature of genome organization (Cremer et al., 1993; Cremer and Cremer, 2010; Heard and Bickmore, 2007). Hence known genomic features are not preserved in $\hat{\mathbf{Z}}_{E,M}^{(r)}$ and $\hat{\mathbf{Z}}_{P,M}^{(r)}$, so these configurations can be disregarded from further analysis.

Visual inspection of the estimated genome configurations from non-metric MDS $\hat{\mathbf{Z}}_{E,NM}^{(r)}$ and $\hat{\mathbf{Z}}_{P,NM}^{(r)}$ (Figure 4.26), reveals the genome taking a spherical shape and chromosomes

occupying individual territories within the sphere, consistent with the chromosome territories feature of genome organization. The estimated genome configurations $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ both look similar and share similar sized $S_3(\hat{\mathbf{Z}})$ (2.14) values, the shape difference POSS$(\mathbf{Y}, \mathbf{X})$ (4.10) was measured for $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. A random POSS$(\mathbf{Y}, \mathbf{X})$ (4.10) obtained by permuting rows of $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$, to serve as a benchmark to measure the POSS$(\mathbf{Y}, \mathbf{X})$ against.

The Table 4.13 shows that $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ are similar in shape and share a similar sized $S_3(\hat{\mathbf{Z}})$ value. In view of this similarity, $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ is used here for further analysis. The stress values in Table 4.13 are poor (Mardia et al., 1979) for both configurations.

| $r$ | $S_3(\hat{\mathbf{Z}}^{(r)}_{E,NM})$ | $S_3(\hat{\mathbf{Z}}^{(r)}_{P,NM})$ | POSS$(\hat{\mathbf{Z}}^{(r)}_{E,NM}, \hat{\mathbf{Z}}^{(r)}_{P,NM})$ | Random POSS$(\hat{\mathbf{Z}}^{(r)}_{E,NM}, \hat{\mathbf{Z}}^{(r)}_{P,NM})$ |
|---|---|---|---|---|
| 3 | 27.3527% | 27.2148% | 0.4087 | 0.9978 |
| 6 | 23.7179% | 23.9368% | 0.2883 | 0.9950 |
| 12 | 22.6204% | 22.1785% | 0.3713 | 0.9917 |
| 24 | 20.2633% | 21.6625% | 0.4360 | 0.9789 |

Table 4.13: Measures of prescaled Procrustes POSS$(\mathbf{X}, \mathbf{Y})$ (4.10) distance between estimated genome configurations $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Column one gives the level of resolution the global Hi-C count matrix has been lowered to (Section 4.8.1). Column two and three give the $S_3(\hat{\mathbf{X}})$ (2.14) values for $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Column four gives the POSS$(\mathbf{X}, \mathbf{Y})$ values between $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Column five gives a mean random POSS$(\mathbf{X}, \mathbf{Y})$ value, found by permuting the rows of $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ and $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. The configurations $\hat{\mathbf{Z}}^{(r)}_{E,NM}$ or $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ are found using by the exponential transform (4.4) or the power transform (4.6) and non-metric MDS with further detail given in Section 4.8.2.

### 4.8.3 Chromosome territories

Lieberman-Aiden et al. (2009) observed that intrachromosomal (within chromosome) contact probability is always larger than interchromosomal (between chromosome) contact probability in $\mathbf{G}^{(1)}$, thus providing evidence for chromosome territories. The

size contrast between interchromosomal and intrachromosomal counts in $\mathbf{G}^{(r)}$ (where $r = 3; 6; 12$ and 24 Mb), will be reversed by the count to distance transform; giving large interchromosomal distances and small intrachromosomal distances. This size contrast in the distances, should influence the non-metric MDS to position the chromosomes in $\hat{\mathbf{Z}}^{(r)}_{.,NM}$ into individual territories. Visually inspecting $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ (in Figure 4.26) reveals that chromosomes do appear to localize into individual territories. A test was developed to see if chromosome territories were preserved during transforming $\mathbf{G}^{(r)}$ into estimated distances, and fitting into three dimensional Euclidean space with non-metric MDS. Let $\mathbf{W}^{(r)} = (w^{(r)}_{a,k})$ be a $23 \times 3$ matrix of the centroids of the human chromosome pairs:

$$w^{(r)}_{a,k} = (n^{(r)}_a)^{-1} \sum_{i=\sum_{c=1}^{a} n^{(r)}_{c-1}+1}^{n^{(r)}_a} z^{(r)}_{i,k} \tag{4.18}$$

is the centroid of chromosome pair "$a$" for "$a$" for $a = 1, \ldots, 22$ and XX, where $n^{(r)}_a$ is the number of points in Chromosome pair "$a$" at resolution $r$ Mb and $n^{(r)}_0 = 0$. The method is as follows.

1. Calculate the mean of the centroids

$$\overline{w}^{(r)}_{.,k} = \frac{1}{23} \sum_{a=1}^{23} w^{(r)}_{a,k}. \tag{4.19}$$

   Then using $\overline{w}^{(r)}_{.,k}$ calculate the variance in the centroids

$$v^{(r)} = \sum_{a=1}^{23} \sum_{k=1}^{3} (w^{(r)}_{a,k} - \overline{w}^{(r)}_{.,k})^2. \tag{4.20}$$

2. Permute the rows of $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ to obtain $\tilde{\mathbf{Z}}^{(r)}_{P,NM}$. Recalculate the matrix of centroids $\tilde{\mathbf{W}}^{(r)}$ using (4.18), ensuring the same ordering of rows are summed over. Recalculate the mean of the centroids $\tilde{\overline{w}}^{(r)}_{.,k}$ (4.19) and the variance in the centroids

$\tilde{v}^{(r)}$ (4.20) for $\tilde{\mathbf{W}}^{(r)}$. Repeat this step 1000 times, collecting a sample of $\tilde{v}^{(r)}$.

3. Count how many $\tilde{v}^{(r)} > v^{(r)}$ and express as a proportion $P^{(r)}$

$$P^{(r)} = \frac{1}{1000}\sum_{i=1}^{1000} I(\tilde{v}_i^{(r)} > v_r). \tag{4.21}$$

The motivation behind this test is the variance in the centroids of $\hat{\mathbf{Z}}_{P,NM}^{(r)}$, should be larger than the variance in the centroids of $\tilde{\mathbf{Z}}_{P,NM}^{(r)}$ (which should cluster by the origin), if the chromosome pairs have been fitted into territories. If the chromosome pair are fitted into territories, their points should colocalize together, and their centroids should be dispersed about the genome. If the chromosome pairs are not fitted into territories, their points will be more dispersed about the genome, and their centroids should colocalize closer to the origin.

The permutation of the rows of $\hat{\mathbf{Z}}_{P,NM}^{(r)}$, will cause the chromosome pairs to exchange points. Then after permutation each of the chromosome pairs points will be dispersed about the genome, causing the centroids to colocalize about the origin. If the chromosome pairs are fitted into territories then in most cases $v^{(r)} > \tilde{v}^{(r)}$, otherwise if they are not fitted into territories then in most cases $v^{(r)} < \tilde{v}^{(r)}$.

The results of the test gave $P^{(r)}$ values of zero on all resolutions, suggesting the chromosome pairs in $\hat{\mathbf{Z}}_{P,NM}^{(r)}$ occupy territories. Therefore the fitting has successfully recovered this aspect of genome organisation.

### 4.8.4 Chromosome clustering

Using the matrix of chromosome centroids $\mathbf{W}^{(r)}$ cluster analysis was performed, to investigate the positioning of the chromosomes as the resolution was lowered. If the clusters consistently contained the same chromosomes, then this showed the non-metric

MDS was recovering aspects from the original genome configuration. If the chromosomes regularly swapped between clusters as resolution was lowered, then the non-metric MDS performed poorly at recovering aspects of the original genome configuration. Regularly swapping between clusters as resolution is lowered is one sign a chromosome is been positioned differently at different resolutions.

First hierarchical clustering was performed, using the average linkage method (same as in Section 4.4). At 1Mb interval resolution (Figure 4.27a) the chromosomes are partitioned into two large clusters. One of these clusters being predominantly composed from large chromosomes and the other small chromosomes. Then on the next level of resolution 3Mb (Figure 4.27b) the dendogram changes, with a number of large chromosomes moving into the smaller chromosome cluster. At the lower levels of resolution (6Mb,12Mb and 24Mb) the shapes of the dendograms (Figures 4.27c,4.27d and 4.27e) change and appear to show aspects of chaining. Chaining is where elements are absorbed into a large cluster in each step of the clustering algorithm, an example of chaining is the dendogram in Figure 4.27c (at 6Mb resolution).

Due to the changing shape of the dendograms it is hard to distinguish how the clusters change as resolution decreases, and the chaining effect might be biasing the analysis. To overcome this k-means clustering (MacQueen et al., 1967) was performed. The k-means clustering partitions the chromosomes into $m$ predefined clusters, where the total distance between the elements of the cluster and its centroid is minimized. Here k-means clustering is used to partition the chromosome centroids into two clusters. The number of swaps between clusters will be counted as resolution decreases, and used to judge if the shape of the estimated genome configuration is consistent as resolution is lowered.

In Table 4.14 resolution decreases from 1Mb to 3Mb there is a single swap between the two clusters; from 3Mb to 6Mb four swaps; from 6Mb to 12Mb another four swaps, and from 12Mb to 24Mb another single swap. Chromosome 9 swaps between the two clusters three times; chromosomes 10, 14 and 18 swap between clusters twice and

chromosome 13 once. This small quantity of swaps between the two clusters and by a few chromosomes, suggest the non-metric MDS is consistently positioning the chromosomes relative to each other as resolution is lowered. Therefore some aspect of the original genome configuration is recovered by non-metric MDS.

| Chromosome | 1 Mb | 3 Mb | 6 Mb | 12 Mb | 24 Mb |
|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 2 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 |
| 3 | 2 | 2 | 2 | 2 | 2 |
| 4 | 2 | 2 | 2 | 2 | 2 |
| 5 | 2 | 2 | 2 | 2 | 2 |
| 6 | 2 | 2 | 2 | 2 | 2 |
| 7 | 2 | 2 | 2 | 2 | 2 |
| 8 | 2 | 2 | 2 | 2 | 2 |
| 9 | 2 | 1 | 2 | 1 | 1 |
| 10 | 1 | 1 | 2 | 1 | 1 |
| 11 | 2 | 2 | 2 | 2 | 2 |
| 12 | 2 | 2 | 2 | 2 | 2 |
| 13 | 2 | 2 | 2 | 2 | 1 |
| 14 | 1 | 1 | 2 | 1 | 1 |
| 15 | 1 | 1 | 1 | 1 | 1 |
| 16 | 1 | 1 | 1 | 1 | 1 |
| 17 | 1 | 1 | 1 | 1 | 1 |
| 18 | 1 | 1 | 2 | 1 | 1 |
| 19 | 1 | 1 | 1 | 1 | 1 |
| 20 | 1 | 1 | 1 | 1 | 1 |
| 21 | 1 | 1 | 1 | 1 | 1 |
| 22 | 1 | 1 | 1 | 1 | 1 |
| X | 2 | 2 | 2 | 2 | 2 |

Table 4.14: Clusters the chromosomes are positioned into, in the estimated genome configuration $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Column one lists the chromosome pairs. Column two to six give the cluster number (1 or 2) the chromosome pair is assigned to, as resolution of $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ decreases. The configuration $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ is found using the power transform (4.6) and non-metric MDS with further detail given in Section 4.8.2. The clustering is found using k-means clustering on the centroids of the chromosome pairs $\mathbf{W}^{(r)}$ (4.18).

### 4.8.5 Radial positioning

The radial positioning of the chromosomes in $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ can be compared with other studies (Tanabe et al., 2002; Boyle et al., 2001; Croft et al., 1999; Rohlf et al., 1980) on chromosome radial positioning to further validate $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. If ordering of chromosome pairs from the origin matches with other studies, then transforming and fitting $\mathbf{G}^{(r)}$ successfully recovered this aspect of $\mathbf{Z}$ in $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Radial positions were calculated using the radial distance $\xi^{(r)}_i$ of each point in $\hat{\mathbf{Z}}^{(r)}_{P,NM}$

$$\xi^{(r)}_i = \left( \sum_{k=1}^{3} z^{(r)^2}_{i,k} \right)^{\frac{1}{2}}. \tag{4.22}$$

Note non-metric MDS centres the fitted configuration at the origin after each iteration, so no value requires subtracting from $z^{(r)}_{i,k}$ in (4.22). Then the mean radial distance for chromosome pair 1 is

$$\overline{\xi}^{(r)}_1 = n_1^{(r)^{-1}} \sum_{i=1}^{n_1^{(r)}} \xi^{(r)}_i, \tag{4.23}$$

and the mean radial distance for each chromosome pair "$a$" for $a = 2, \ldots, 22$ and XX is

$$\overline{\xi}^{(r)}_a = n_a^{(r)^{-1}} \sum_{i=n_1^{(r)}+\ldots+n_{a-1}^{(r)}+1}^{n_a^{(r)}} \xi^{(r)}_i. \tag{4.24}$$

The chromosomes were positioned in increasing order of $\overline{\xi}^{(r)}_a$ for for $a = a, \ldots, 22$ and XX, this is displayed in Table 4.15. In Table 4.15 the radial ordering of the chromosome pairs from the origin, is different on each resolution used. This lack of consitency in the radial ordering confounds the interpretation of the data. Chromosome $19$ is consistently fitted on the periphery of $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ while chromosome $18$ is consistently fitted on the interior of $\hat{\mathbf{Z}}^{(r)}_{P,NM}$, contradicting Tanabe et al. (2002) and Croft et al. (1999) who observe

| Chromosome | 3 Mb | 6 Mb | 12 Mb | 24 Mb |
|---|---|---|---|---|
| 1 | 12 | 12 | 12 | 9 |
| 2 | 16 | 16 | 15 | 15 |
| 3 | 10 | 8 | 8 | 8 |
| 4 | 20 | 20 | 20 | 21 |
| 5 | 6 | 11 | 11 | 12 |
| 6 | 19 | 14 | 16 | 18 |
| 7 | 15 | 15 | 17 | 14 |
| 8 | 9 | 10 | 9 | 10 |
| 9 | 7 | 6 | 4 | 6 |
| 10 | 5 | 5 | 6 | 4 |
| 11 | 1 | 1 | 2 | 2 |
| 12 | 11 | 9 | 10 | 11 |
| 13 | 17 | 18 | 19 | 20 |
| 14 | 8 | 7 | 7 | 7 |
| 15 | 3 | 3 | 3 | 3 |
| 16 | 13 | 13 | 14 | 16 |
| 17 | 18 | 17 | 13 | 13 |
| 18 | 4 | 4 | 5 | 5 |
| 19 | 23 | 23 | 23 | 23 |
| 20 | 2 | 2 | 1 | 1 |
| 21 | 21 | 21 | 21 | 19 |
| 22 | 14 | 19 | 18 | 17 |
| X | 22 | 22 | 22 | 22 |

Table 4.15: Radial ordering from the origin of the chromosome pairs, in the estimated genome configuration $\hat{\mathbf{Z}}^{(r)}_{P,NM}$. Column one lists the chromosome pairs. Column two to five give the radial ordering the chromosome pair takes from the origin, as resolution of $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ decreases. The configuration $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ is found using the power transform (4.6) and non-metric MDS with further detail given in Section 4.8.2. The distance of the chromosome pairs from the origin is found using (4.22), (4.23) and (4.24).

chromosome $18$ localizes near the nuclear periphery while chromosome $19$ localizes near the nuclear interior. Studies on the ordering of the chromosomes in male human lymphoblast nucleus Boyle et al. (2001) found chromosomes $1, 16, 17, 19$ and $22$ localize near the nuclear interior, chromosomes $5, 6, 10, 14, 15$ and $21$ localize in an intermediate region and chromosomes $2, 3, 4, 7, 8, 9, 11, 12, 13, 18$ X and Y localize near the nuclear periphery. The ordering of chromosomes in $\hat{\mathbf{Z}}^{(r)}_{P,NM}$ displays almost an opposite ordering to Boyle et al. (2001) with chromosome which should be located near the interior fitted to the periphery and vice-versa for the chromosomes which should be located near the periphery. Lowering resolution, transforming and fitting have failed to recover the radial positions of the chromosomes in the genome.

### 4.8.6 Discussion

Lowering the resolution when scaling the genome is useful to lower noise and aid interpretation. The $\hat{\mathbf{Z}}^{(r)}$ do not recover known features of the genome and can be deemed poor estimates for the genome. The poor recovery of $\hat{\mathbf{Z}}^{(r)}$ could be driven by the fact that the interchromosomal counts are very small in comparison to the intrachromosomal counts, even after lowering resolution. Another factor contributing to the poor $\hat{\mathbf{Z}}^{(r)}$ could be that chromosome pairs are been fitted into Euclidean space, not the individual chromosomes. This could be a problem as for example, the copies of chromosome one could be reflections of each other in the xy-axis (Figure 1.1). Then the fitted chromosome pair could take the average of the copies location and fit the chromosome pair close to the origin.

## 4.9 Conclusion

Transforming and fitting $\mathbf{M}$ from chromosome 14 has recovered some interesting features in $\hat{\mathbf{X}}$, which could correspond to features of $\mathbf{X}$ but also recovers horseshoe shapes which could be from horseshoe effect (Section 2.4). Applying cluster analysis to $\hat{\mathbf{D}}_{E,M}$ has uncovered features which correspond to the fractal globule model (Section 1.2.6) in $\hat{\mathbf{X}}_{E,M}$. Inspecting the count frequency in $\mathbf{M}$ has uncovered partitioning which correspond to the two component model in $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,M}$. This clustering and partitioning shows that some local structure in $\hat{\mathbf{X}}$ is preserved. Scaling symmetric submatrices from $\mathbf{M}$ to observe local structure without the influence of the horseshoe effect, only serves to make medium distances the new large distances and the horseshoe effect repeats itself in the fitted subregion.

Investigating the relationship between $m_{i,j}$ and $\nu_{i,j}$, we observe a plateauing of the $m_{i,j}$ at $\nu_{i,j} \geq 15$Mb. This plateauing confounds estimating $d_{i,j}$ at this range, as inaccuracies in $\tilde{d}_{i,j}$ at $\nu_{i,j} \geq 15$Mb will contribute to the horseshoe effect in $\hat{\mathbf{X}}$.

Future recovery of $\mathbf{X}$ requires improvement in $\tilde{d}_{i,j}$ estimation through investigating alteration of score and transform functions. Score functions which are distance-based instead of count-based may be preferable, as the reverse transform function can magnify changes in small distances to produce very large counts, which can be detrimental to $\hat{\mathbf{X}}$.

Recovery of genome configuration preserves the chromosome territories but does not preserve ordering of the chromosomes from the origin. The contrast between intrachromosomal and interchromosomal counts can explain the preservation of territories. The low interchromosomal counts provide little information on chromosome positioning even when resolution is lowered. Genome recovery would require more interchromosomal information to be available to be successful.

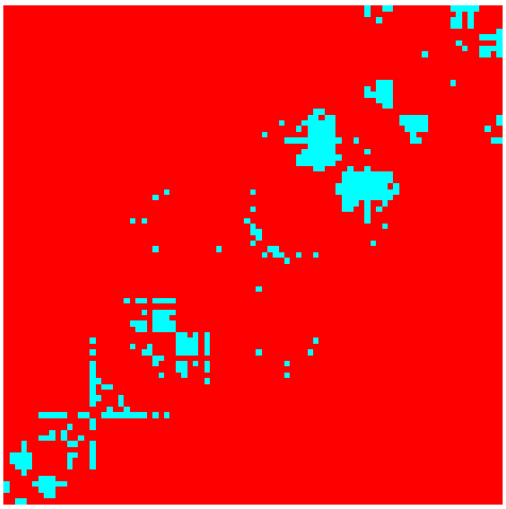


Figure 4.24: Locations of filtered counts from Chromosome 14's filtered Hi-C counts matrix $\tilde{\mathbf{M}}$. The blue squares ■ denote the filtered counts. The red squares ■ denote the unfiltered counts. The filtered count matrix $\tilde{\mathbf{M}}$ is found by applying the filtering process in Section 4.7.2 with threshold $\tau = 0.05$, to Chromosome 14's Hi-C count matrix.

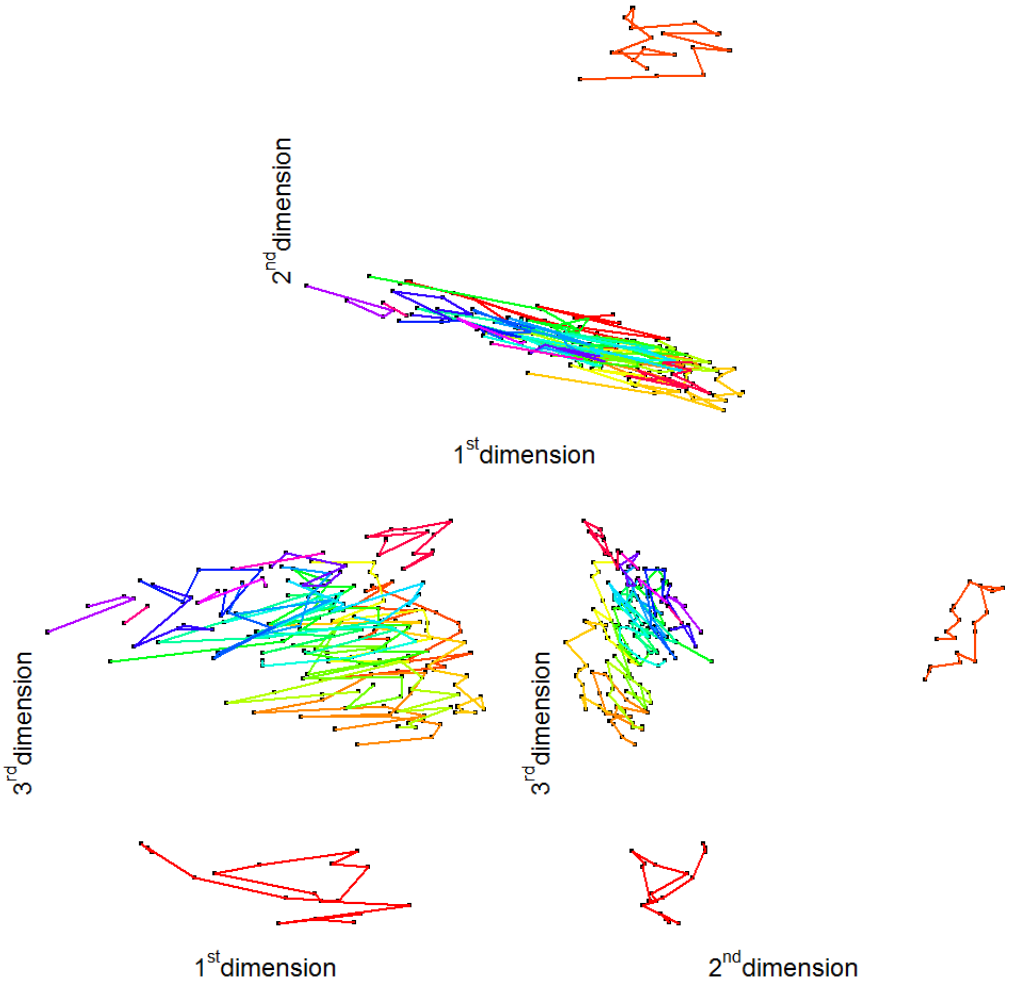



Figure 4.25: Perspectives of the estimated genome configuration $\hat{\mathbf{Z}}_{P,M}^{(12)}$, found using the power transform (4.6) and fitting into three dimensional Euclidean space with metric MDS. Further details on finding $\hat{\mathbf{Z}}_{P,M}^{(12)}$ can be found in Section 4.8.2. The colours define which chromosome pairs the points of $\hat{\mathbf{Z}}_{P,M}^{(12)}$ belong to.

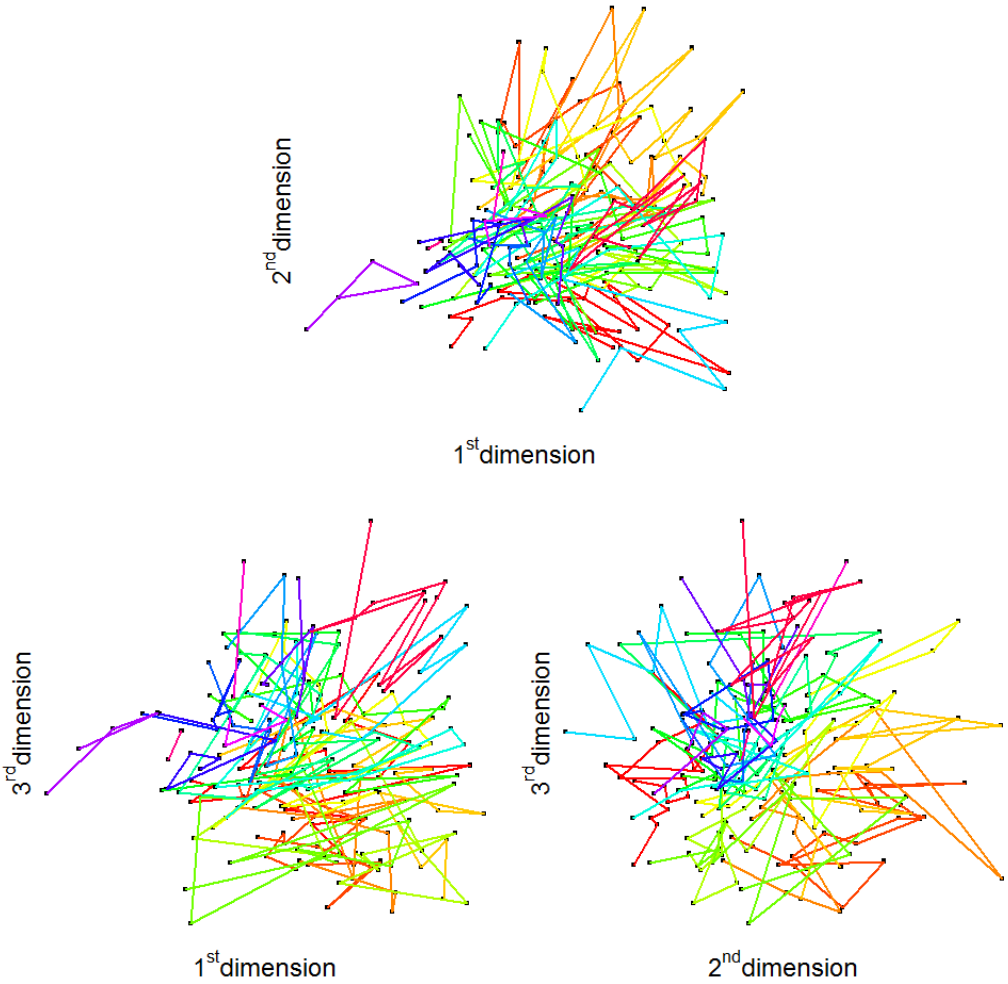
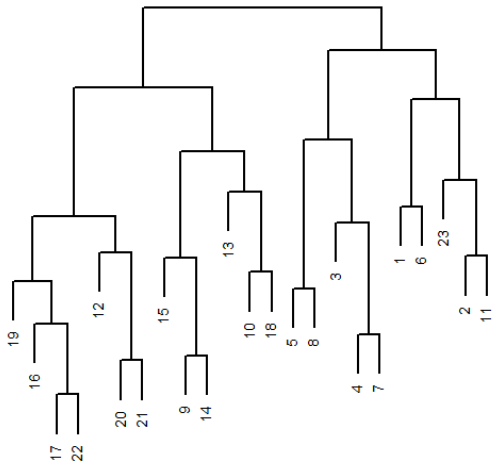
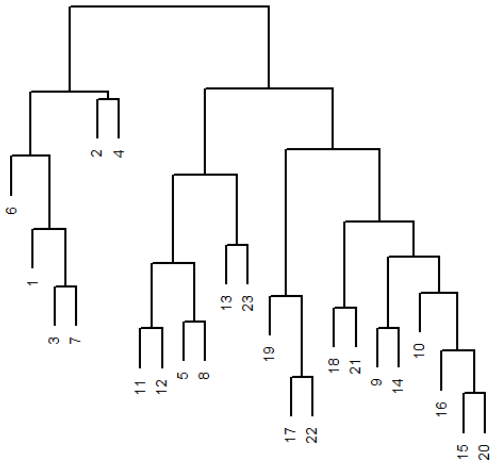


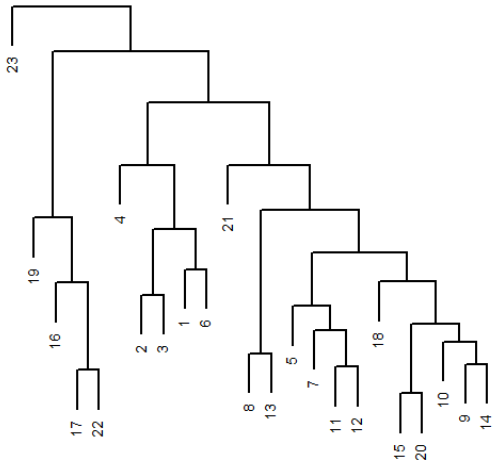

Figure 4.26: Perspectives of the estimated genome configuration $\hat{\mathbf{Z}}^{(12)}_{P,NM}$, found using the power transform (4.6) and fitting into three dimensional Euclidean space with non-metric MDS. Further details on finding $\hat{\mathbf{Z}}^{(12)}_{P,NM}$ can be found in Section 4.8.2. The colours define which chromosome pairs the points of $\hat{\mathbf{Z}}^{(12)}_{P,NM}$ belong to.



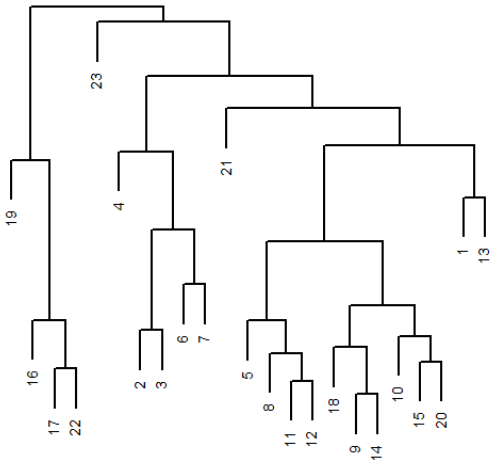

(a) 1Mb resolution



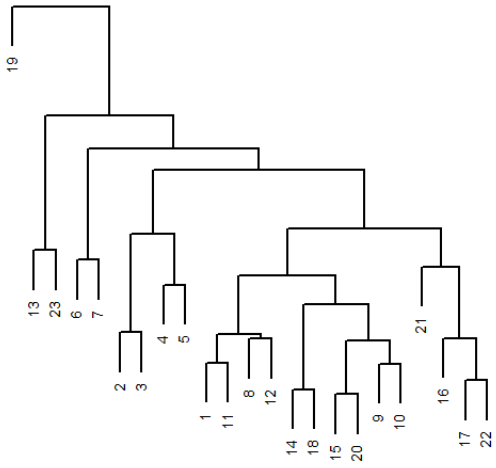

(b) 3Mb resolution




(c) 6Mb resolution




(d) 12Mb resolution




(e) 24Mb resolution

Figure 4.27: Dendograms displaying how the chromosome are positioned relative to each other, in the estimated genome configuration. The estimated genome configuration is found using the power transform and non-metric MDS at resolution $r$Mb, which gives $\mathbf{Z}^{(r)}_{P,NM}$. Then the 23×3 matrix of chromosome centroids $\mathbf{W}^{(r)}$ is found using (4.18). Finally average linkage cluster analysis is performed to give the above dendograms.

# Chapter 5

# Model-based approach

The preceding chapters used the tools of a count-to-distance transform and multidimensional scaling (MDS) to recover an estimated chromosome configuration (or estimated genome configuration) from a chromosome count matrix (or global count matrix). The true chromosome configuration and the true relationship between counts and distances are unknown. These unknowns mean assumptions about the count to distance relationship have to be made to produce a suitable transform function, to obtain estimated distances. The estimated chromosome configuration obtained from fitting the estimated distances into three dimensional Euclidean space, is assumed to be close to the true chromosome configuration. Although the counts in the chromosome count matrix contain noise and little is known on how this noise affects the estimated distances and the estimated chromosome configuration. To find how the noise in the counts affects the estimated distances and fitted configuration, a model-based approach was adopted (MBA). The MBA used known configurations to produce matrices of mean counts, these were then perturbed to give perturbed count matrices. The perturbed count matrices were fitted into Euclidean space in a similar way to the chromosome count matrices. Knowing the original configuration and the true relationship between counts and distances provides a platform for investigating the affects of noise in the counts.

The MBA takes an initial configuration $\mathbf{X} = (x_{i,k})$ of $n$ points sitting in $p$ dimensional space; extracts a Euclidean distance matrix $\mathbf{D} = (d_{i,j})$ from $\mathbf{X}$ using (2.2) and, then inverts the count to distance transform function to obtain a matrix of mean counts $\mathbf{U} = (\mu_{i,j})$. Using $\mathbf{U}$ and a chosen level of dispersion $\rho$, a matrix of perturbed counts $\mathbf{M} = (m_{i,j})$ is then simulated, and transformed into perturbed distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$. The perturbed distances are then fitted into $p$ dimensional space using MDS to obtain a fitted configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$. Finally, MDS internal measures of fit and measures comparing the initial and fitted configuration are collected. Similar approaches were used by Sibson et al. (1981), where non-independent perturbation was added to distances, $\hat{\mathbf{X}}$ was recovered using metric; non-metric or least squares MDS and compared with $\mathbf{X}$ using Procrustes statistics. Informally the MBA can be thought of as walking up a mountain during the day then finding a return path at night, finally comparing start and end points of the walk. The MBA gives control over the initial configuration; how the counts are

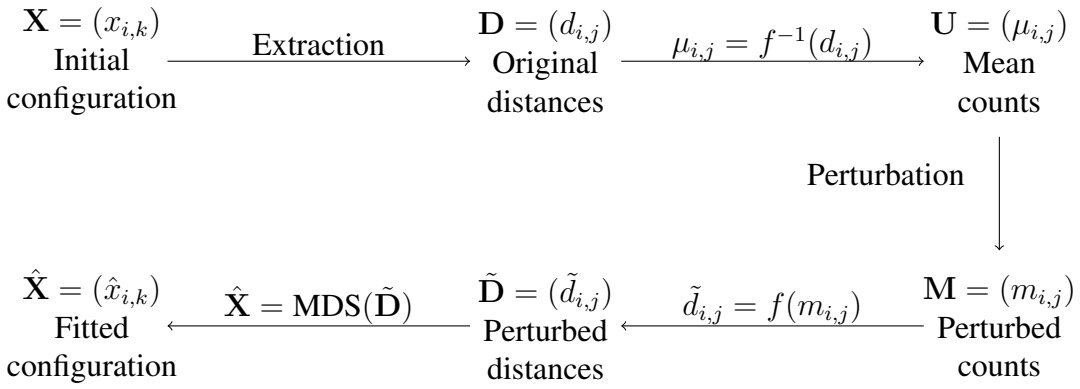



Figure 5.1: Schematic summarizing the model based approach. Where $f(\ldots)$ is the count to distance transform function (4.4) or (4.6); $f^{-1}(\ldots)$ is the inverse transform function (4.5) or (4.7), and MDS is multidimensional scaling.

transformed into distances; how the counts are perturbed and the method of MDS used, this control allows adjustments to be made to see how they affect the scaling process and the fitted configuration. Control over the initial configuration provides a platform for the MBA. Changing the number of dimensions in the initial configuration and its complexity

all affect the MBA. For example changing the number of dimensions can decrease count size, by increasing the Euclidean distances. Changing the complexity of the configuration can change the size and distribution of counts within $\mathbf{U}$. For example in a straight line with equally spaced points, $\mathbf{U}$ is Toeplitz with count size decreasing on the subdiagonals away from the main diagonal; for a circle with equally spaced points, $\mathbf{U}$ is also Toeplitz with count size decreasing then increasing on the subdiagonals away from the main diagonal. Control over the count to distance transform allows the transform used to transform perturbed counts into perturbed distances to be the same transform which the inverse transformed the distances into mean counts. It also provides control over the size of the mean counts. Control over the level of dispersion when simulating the perturbed counts affects how much information on the initial configuration is retained in the perturbed counts. Once the controls have been chosen $\tilde{\mathbf{D}}$ can be fitted into Euclidean space using metric or non-metric MDS to find which MDS method performs better. Adjusting these controls and monitoring how they affect the fitted configuration will give insights to help improve chromosome configuration estimation.

## 5.1 Constructing the MBA

### 5.1.1 Initial configuration

Four initial configurations were used in the MBA: a straight line; a parabola; a semi-circle and a circle. These shapes were used as initial configurations because of their simplicity and low dimensionality allowed clear visual comparison between the initial and fitted configurations, and inspection of the noise in the spare (second or third) dimensions. Each configuration consisted of $n = 100$ points: this was considered a sensible amount, too many points made visual comparison difficult and too few points provided too little information to obtain useful results. To make the data more comparable each shape was

scaled to have a maximum distance of one unit ($d_{max} = 1$). The initial configurations can be written as $\mathbf{X} = (\underline{x}_{(1)}, \underline{x}_{(2)})$ where $\underline{x}_{(k)} = (x_{i,k})$ are the vectors of coordinates for the points $i = 1, \ldots, 100$ in dimensions $k = 1, 2$.

**Straight line**

$$x_{i,1} = -\frac{1}{2} + \frac{(i-1)}{99} \text{ and } x_{i,2} = 0.$$

**Parabola**

$$x_{i,1} = -\frac{1}{2} + \frac{(i-1)}{99} \text{ and } x_{i,2} \left( -\frac{1}{2} + \frac{(i-1)}{99} \right)^2.$$

**Semi-circle**

$$x_{i,1} = \frac{1}{2}\cos\left(\frac{\pi}{99}(i-1)\right) \text{ and } x_{i,2} = \frac{1}{2}\sin\left(\frac{\pi}{99}(i-1)\right).$$

**Circle**

$$x_{i,1} = \frac{1}{2}\cos\left(\frac{2\pi}{100}(i-1)\right) \text{ and } x_{i,2} = \frac{1}{2}\sin\left(\frac{2\pi}{100}(i-1)\right).$$

### 5.1.2 Count to distance transform

In the MBA the transform used to transform the original distances into mean counts is the same as the transform used to transform the perturbed counts into perturbed distances. This allows the perturbed distances to hold some resemblance to the original distances and avoids the need to estimate the transforms parameters. The exponential and power transforms were investigated independently in the MBA using different parameters to alter

mean count size. The data generated from the two transforms cannot be compared due to the different nature of the transforms.

### Exponential transform

The exponential transform and its inverse described in (4.4) and (4.5) are respectively,

$$d_{i,j} = e^{-\alpha\mu_{i,j}} \text{ and } \mu_{i,j} = -\frac{1}{\alpha}\log(d_{i,j}).$$

The exponential transform was investigated on four levels of $\alpha$, setting $\alpha = 0.1; 0.01; 0.001$ and $0.0001$. As $\alpha$ decreases the mean count size increases.

### Power transform

The power transform and its inverse described in (4.6) and (4.7) are respectively,

$$d_{i,j} = (b_0\mu_{i,j})^{\beta} \text{ and } \mu_{i,j} = \frac{d_{i,j}^{\frac{1}{\beta}}}{b_0}.$$

The power transform was investigated at four levels of $b_0$, setting $b_0 = 0.1; 0.01; 0.001$ and $0.0001$. As $b_0$ decreases the mean count size increases. The $\beta$ parameter was held constant at $\beta = -0.5$ for each level of $b_0$. The $m_{\text{min}} = 1$ and $m_{\text{min}} = 2$ adjustments were applied when using the power transform providing two sets of results.

## 5.1.3 The structure of the noise

The chromosome contact matrix already contains noise embedded within it, whereas in the MBA the noise is introduced into the $\mu_{i,j}$ to obtain the $m_{i,j}$. The distributions best suited for introducing noise into $\mu_{i,j}$ are the Poisson or the negative binomial distributions,

depending on the required level of dispersion $\rho$. The dispersion measures how many times larger the variance is with respect to the expectation

$$\rho = \frac{\text{var}(m_{i,j})}{E(m_{i,j})}. \tag{5.1}$$

## Poisson distribution

The Poisson distribution is used to generate $m_{i,j}$ when $\rho = 1$

$$m_{i,j} \sim \text{Poisson}(\mu_{i,j}), \tag{5.2}$$

and so $\text{E}(m_{i,j}) = \mu_{i,j}$ and $\text{var}(m_{i,j}) = \mu_{i,j}$.

## Negative binomial distribution

The negative binomial distribution is used to generate $m_{i,j}$ when $\rho > 1$ (over-dispersed Poisson),

$$m_{i,j} \sim \text{NB}(r, l), \tag{5.3}$$

where $r = \frac{\mu_{i,j}}{\rho - 1}$ and $l = \frac{1}{\rho}$, providing $E(m_{i,j}) = \mu_{i,j}$ and $\text{var}(m_{i,j}) = \rho\mu_{i,j}$. The levels of dispersion used were $\rho = 1; 2; 4$ and $8$.

The size of the $\mu_{i,j}$ and $\rho$ play a role in how much perturbation is translated into $m_{i,j}$, which can be explained using the coefficient of variation $C_v(\mu_{i,j}, \rho)$,

$$\text{C}_v(\mu_{i,j}, \rho) = \frac{\sqrt{\text{var}(m_{i,j})}}{\text{E}(m_{i,j})} = \sqrt{\frac{\rho}{\mu_{i,j}}} \tag{5.4}$$

The larger $C_v(\mu_{i,j}, \rho)$ is the more noise (perturbation) is translated into $m_{i,j}$. For example, for $\mu_{i,j} = 100$ and $\rho = 1$ we have $C_v(\mu_{i,j}, \rho) = 0.1$ and a value $m_{i,j} = 105$ would

not be unusual. For $\mu_{i,j} = 10$ and $\rho = 1$, we have $C_v(\mu_{i,j}, \rho) = 0.316$ and a value $m_{i,j} = 15$ would not be unusual. Perturbation adds the same quantity in each case, but in the first case perturbation increases count size by $5\%$ and the second $50\%$. The $50\%$ increase in the second case could cause a much larger decrease in distance than the first. The size of $\mu_{i,j}$ and $\rho$ determine the amount of perturbation translated into $m_{i,j}$, similar to the intuition that large counts hold more information on distances than smaller counts. Variation in $C_v(\mu_{i,j}, \rho)$ is illustrated in Table 5.1 for each transform, using $\mu_{i,j}$ corresponding to $d_{i,j} = 0.5$. Both tables display a decrease in $C_v(\mu_{i,j}, \rho)$ as $\mu_{i,j}$ increases (moving down the tables) and a increase in $C_v(\mu_{i,j}, \rho)$ as $\rho$ increases (moving left to right in the table).

**Exponential transform**

| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
|---|---|---|---|---|
| 0.1 | 0.3798 | 0.5372 | 0.7597 | 1.0743 |
| 0.01 | 0.1201 | 0.1699 | 0.2402 | 0.3397 |
| 0.001 | 0.0380 | 0.0537 | 0.0760 | 0.1074 |
| 0.0001 | 0.0120 | 0.0170 | 0.0240 | 0.0340 |

**Power transform**

| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
|---|---|---|---|---|
| 0.1 | 0.1581 | 0.2236 | 0.3162 | 0.4472 |
| 0.01 | 0.0500 | 0.0707 | 0.1000 | 0.1414 |
| 0.001 | 0.0158 | 0.0224 | 0.0316 | 0.0447 |
| 0.0001 | 0.0050 | 0.0071 | 0.0100 | 0.0141 |

Table 5.1: Illustration of how altering transform parameters and dispersion affect the coefficient of variation $C_v(\mu_{i,j}, \rho)$ (5.4). Top table: how $C_v(\mu_{i,j}, \rho)$ changes for the exponential transform (4.4) when $\alpha$ decreases and the dispersion $\rho$ increases. Bottom table: how $C_v(\mu_{i,j}, \rho)$ changes for the power transform (4.6) when $b_0$ decreases and $\rho$ increases. The $C_v(\mu_{i,j}, \rho)$ are found by finding the mean count $\mu_{i,j}$ for a distance $d_{i,j} = 0.5$ by using either (4.5) with $\alpha$ or (4.7) with $b_0$ and $\beta = -0.5$. The inputting $\mu_{i,j}$ with $\rho$ into (5.4).

### 5.1.4 Multidimensional scaling

A critical step of the MBA is the recovery of a fitted configuration $\hat{\mathbf{X}}$ from the perturbed distance matrix $\tilde{\mathbf{D}}$. In the MBA, either metric or non-metric multidimensional scaling (MDS) was used to recover $\hat{\mathbf{X}}$ and accompanying MDS measures of fit $\theta_{1:p}$ (2.12) or $S_p(\hat{\mathbf{X}})$ (2.14) were used (see Chapter 2); where $p$ is the number of dimensions in $\mathbf{X}$.

### 5.1.5 Assessing the fit

The initial and fitted configurations $\mathbf{X}$ and $\hat{\mathbf{X}}$ are compared by measuring shape difference, expansion in size and through visual comparison. The shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ measures how well $\mathbf{X}$ has been recovered in $\hat{\mathbf{X}}$. This is the most versatile statistic as it can be applied to $\hat{\mathbf{X}}$ from either MDS method. The size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$ measures how much $\hat{\mathbf{X}}$ has expanded with respect to $\mathbf{X}$. Visual comparison between $\mathbf{X}$ and $\hat{\mathbf{X}}$ is used to assess where the discrepancies lie and how noise is distributed in $\hat{\mathbf{X}}$. The MDS's performance is measured using $\theta_{1:p}$ (2.12) or $S_p(\hat{\mathbf{X}})$ (2.14) for metric or non-metric MDS respectively.

#### Shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$

The shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ is based on the Procrustes shape distance between $\mathbf{X}$ and $\hat{\mathbf{X}}$ (2.16) normalized by $\mathbf{X}$'s size (Mardia et al., 1979),

$$P(\mathbf{X}, \hat{\mathbf{X}}) = \left( \frac{\text{OSS}(\mathbf{X}, \hat{\mathbf{X}})}{\text{tr}(\mathbf{X}^T\mathbf{X})} \right)^{\frac{1}{2}}, \tag{5.5}$$

where $\hat{\mathbf{X}}$ is fitted into the same number of dimensions as $\mathbf{X}$. A value of $P(\mathbf{X}, \hat{\mathbf{X}}) = 0$ indicates identical shapes and as $P(\mathbf{X}, \hat{\mathbf{X}})$ increases $\mathbf{X}$ and $\hat{\mathbf{X}}$ grow more dissimilar.

### Size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$

$G(\mathbf{X}, \hat{\mathbf{X}})$ is based on the ratio of sum of the squared distance from the origin of the points in $\mathbf{X}$ and $\hat{\mathbf{X}}$.

$$G(\mathbf{X}, \hat{\mathbf{X}}) = \left( \frac{\sum_{i=1}^{n} \sum_{k=1}^{p} \hat{x}_{i,k}^2}{\sum_{i=1}^{n} \sum_{k=1}^{p} x_{i,k}^2} \right)^{\frac{1}{2}}. \tag{5.6}$$

If $G(\mathbf{X}, \hat{\mathbf{X}}) > 1$ then $\hat{\mathbf{X}}$ is larger than $\mathbf{X}$. If $G(\mathbf{X}, \hat{\mathbf{X}}) = 1$ the the two configurations are of equal size and if $G(\mathbf{X}, \hat{\mathbf{X}}) < 1$ then $\hat{\mathbf{X}}$ is smaller than $\mathbf{X}$. The measure $G(\mathbf{X}, \hat{\mathbf{X}})$ can only be applied to metric MDS output, as scale is not preserved by non-metric MDS.

### Performance of the MDS

The measures $\theta_{1:p}$ and $S_p(\hat{\mathbf{X}})$ measure the performance of metric or non-metric MDS respectively. The values of $\theta_{1:p}$ should ideally maximized and $S_p(\hat{\mathbf{X}})$ should ideally minimized.

### Visual comparison

Information which cannot be summarized through $P(\mathbf{X}, \hat{\mathbf{X}})$ and $G(\mathbf{X}, \hat{\mathbf{X}})$ might be observed through visual comparison. Visual comparison plots $\mathbf{X}$ and $\hat{\mathbf{X}}$ together (where $\hat{\mathbf{X}}$ is mapped onto $\mathbf{X}$ using a procedure similar to that outlined for OSS$(\mathbf{X}, \hat{\mathbf{X}})$ but with $p = 3$), to locate discrepancies and gain insight into how noise is distributed within $\hat{\mathbf{X}}$.

## 5.1.6 Running the model-based approach

A single run of the MBA provides a single set of MDS performance statistics; a single $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) value and a single $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) value. Repeating the MBA with the same

conditions several times (1000), several values of MDS performance statistics; $P(\mathbf{X}, \hat{\mathbf{X}})$ and $G(\mathbf{X}, \hat{\mathbf{X}})$ values can be obtained. The mean of these values can be found to give more robust MDS performance statistics; $P(\mathbf{X}, \hat{\mathbf{X}})$ and $G(\mathbf{X}, \hat{\mathbf{X}})$ values. The combination of shape, transform function, transform function parameter, $m_{\text{min}}$ adjustment and MDS method gives 384 distinct ways of running the MBA, providing a rich data set.

## 5.2 Simulation results

There are 384 distinct ways of running the MBA, providing a broad range of scenarios to analyse. The simulation results on the MDS performance are analysed first, followed by the shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) and finally the size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6). The simulation results for the semi-circle are displayed in the plots and the simulation results for the other shapes can be found in the Appendix Section B. Comments made incorporate results from all the shapes.

### MDS performance

The trend in the performance of the MDS is the performance improves as coefficient of variation $C_v(\mu_{i,j}, \rho)$ (5.4) decreases (when large mean counts are used or lower dispersion). This trend is observed in each shape, in both transform functions and both MDS methods. The trend can be explained as there is less perturbation in $\tilde{\mathbf{D}}$ as $C_v(\mu_{i,j}, \rho)$ decreases, which makes it easier for the MDS to find a fitted configuration, resembling the initial configuration.

When using metric MDS for either the exponential transform (4.4) or power transform (4.6) when $C_v(\mu_{i,j}, \rho)$ is large, less information $\theta_{1:p}$ (2.12) is projected into the first $p$ dimensions. This suggests the perturbation is producing large spurious eigenvalues which distribute information from the noise into additional dimensions. When using

non-metric MDS for either the exponential or power transform when $C_v(\mu_{i,j}, \rho)$ is large, the $S_p(\hat{\mathbf{X}})$ values are large. Suggesting the perturbation has an effect of permuting the ordering of the distances, which is inflating the $S_p(\hat{\mathbf{X}})$ values. Only on the lowest level of $C_v(\mu_{i,j}, \rho)$ are the $S_p(\hat{\mathbf{X}})$ values better than fair (Mardia et al., 1979) for the exponential transform. The $S_p(\hat{\mathbf{X}})$ are more robust to perturbation when using the power transform and only on the highest level of $C_v(\mu_{i,j}, \rho)$ are the $S_p(\hat{\mathbf{X}})$ values unfair. The $m_{\text{min}}$ adjustment applied to the power transform, has only produced an improvement to the $\theta_{1:p}$ statistic on the $b_0$ level. The requirement for $m_{\text{min}}$ adjustment diminishes as $b_0$ increases, as the $\mu_{i,j}$ become larger and the probability of generating a $m_{i,j} = 1, 2$ diminishes.

## Shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$

Across all shapes and both transforms the trend observed is that $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) improves as $C_v(\mu_{i,j}, \rho)$ decreases. Hence in most cases the fitted configuration is larger than the original configuration. When using the exponential transform (4.4) the MDS method which gives best $P(\mathbf{X}, \hat{\mathbf{X}})$ values vary from shape to shape, but when $C_v(\mu_{i,j}, \rho)$ is small the difference is negliable. When using the power transform (4.6), non-metric MDS appears to give a better value for $P(\mathbf{X}, \hat{\mathbf{X}})$ when $C_v(\mu_{i,j}, \rho)$ is large, and metric MDS gives the better values when $C_v(\mu_{i,j}, \rho)$ is small. The $m_{\text{min}} = 2$ adjustment denoted by the dashed lines in Figure 5.3b, only provides an better $P(\mathbf{X}, \hat{\mathbf{X}})$ value at $b_0 = 0.1$ when using metric MDS, although the size of the improvement is substantial.

## Size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$

The trend observed across all four shapes and both transform functions with the exception of the circle using the exponential transform, is $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) is larger than 1 and decreases to 1 as the $C_v(\mu_{i,j}, \rho)$ decreases. For the circle $G(\mathbf{X}, \hat{\mathbf{X}})$ increases to 1 as $C_v(\mu_{i,j}, \rho)$ decreases. This size increase for the circle could be due to opposing points

on the circle sharing $d_{i,j} = 1$ give a $\mu_{i,j} = 0$ which generates a $m_{i,j} = 0$ so the distance between opposing points remains unperturbed, these unperturbed distances preventing the circle expanding. The $m_{\text{min}} = 2$ adjustment denoted by the dashed lines in Figure 5.4b appears to improve $G(\mathbf{X}, \hat{\mathbf{X}})$ at $b_0 = 0.1$.

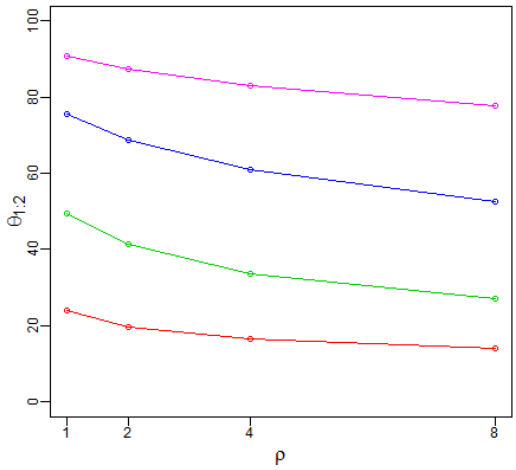
### 5.2.1 Simulations summary

The trend observed across the simulations is the recovery of the fitted configuration improves as $C_v(\mu_{i,j}, \rho)$ decreases. This improvement can be expected as at lower $C_v(\mu_{i,j}, \rho)$ less perturbation is present in the perturbed distances. The shape difference statistic is the strongest indicator, at which MDS method performed better at different levels of $C_v(\mu_{i,j}, \rho)$. When using the exponential transform, the preference for metric or non-metric MDS varied from shape to shape. When using the power transform, non-metric MDS appeared to perform better when $C_v(\mu_{i,j}, \rho)$ was large and metric MDS performed better when $C_v(\mu_{i,j}, \rho)$ was smaller. The size expansion statistic indicated that $\hat{\mathbf{X}}$ were becoming larger than the $\mathbf{X}$. This suggests some mechanism in the perturbation is increasing the size of perturbed distances, and this increase is being translated into $\hat{\mathbf{X}}$. The poorest $\hat{\mathbf{X}}$ are generated using the power transform and metric MDS with the parameters $b_0 = 0.1$ and $m_{\text{min}} = 1$ here the $P(\mathbf{X}, \hat{\mathbf{X}})$ and $G(\mathbf{X}, \hat{\mathbf{X}})$ values appear more than twice the size of the next level of $b_0$ (at $b_0 = 0.01$). Although the $m_{\text{min}} = 2$ adjustment does provide some improvement to $\hat{\mathbf{X}}$. Therefore simulations give evidence that replacing the $m_{i,j} = 0$ or 1 with $m_{i,j} = 2$ does provide some improvement to $\hat{\mathbf{X}}$.

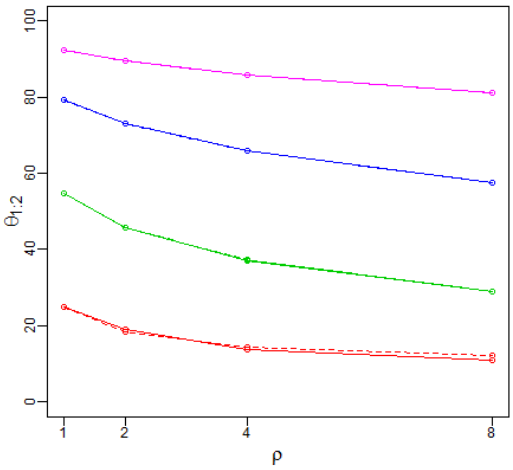
### 5.2.2 Visual comparison

Visual comparison was used for a semi-circle generated under conditions which give the poorest fitting $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). When the $\hat{\mathbf{X}}$ from metric MDS is displayed, the fitted

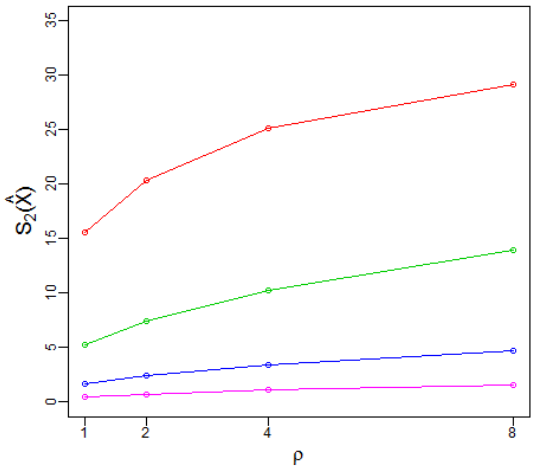
(a) Exponential transform & metric MDS.




(b) Power transform & metric MDS.




(c) Exponential transform & non-metric MDS.




(d) Power transform & non-metric MDS.

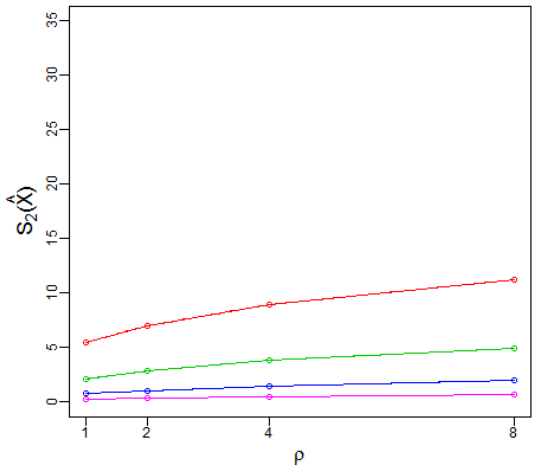
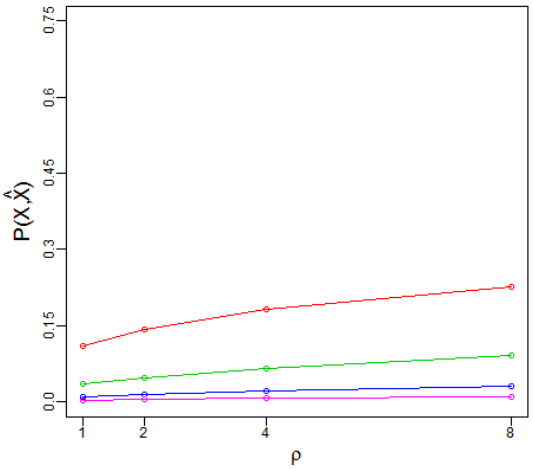
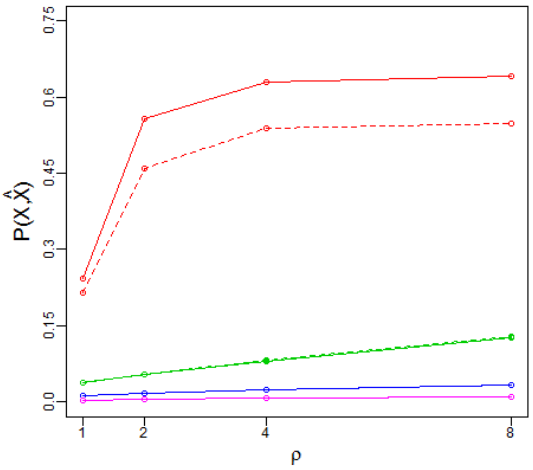



Figure 5.2: MDS performance statistics $\theta_{1:2}$ (2.12) or $S_2(\hat{\mathbf{X}})$ (2.14) from the MBA simulations for a semi-circle. Column one: perturbed distances $\tilde{\mathbf{D}}$ are generated using the exponential transform (4.4). Column two: $\tilde{\mathbf{D}}$ are generated using the power transform (4.6). Row one: $\theta_{1:2}$ values from fitting $\tilde{\mathbf{D}}$ into two dimensional Euclidean space with metric MDS. Row two: $S_2(\hat{\mathbf{X}})$ values from fitting $\tilde{\mathbf{D}}$ into two dimensional Euclidean space with non-metric MDS. The red lines for $\alpha = 0.1$ (4.4) or $b_0 = 0.1$ (4.6); the green lines for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines for $\alpha = 0.0001$ or $b_0 = 0.0001$. In the power transform figure the solid lines signify the $m_{\text{min}} = 1$ adjustment has been applied, and the dashed lines signify the $m_{\text{min}} = 2$ adjustment has been applied.

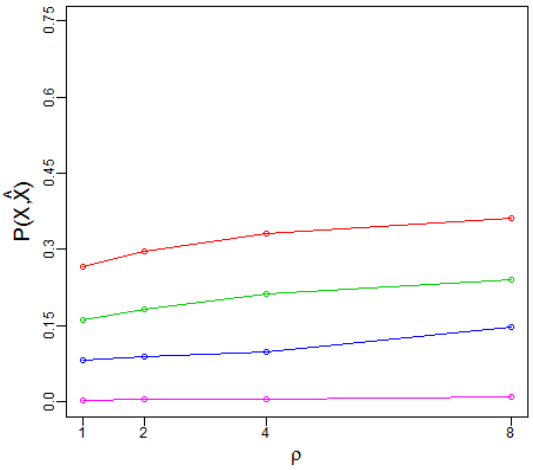
(a) Exponential transform & metric MDS.




(b) Power transform & metric MDS.




(c) Exponential transform & non-metric MDS.




(d) Power transform & non-metric MDS.

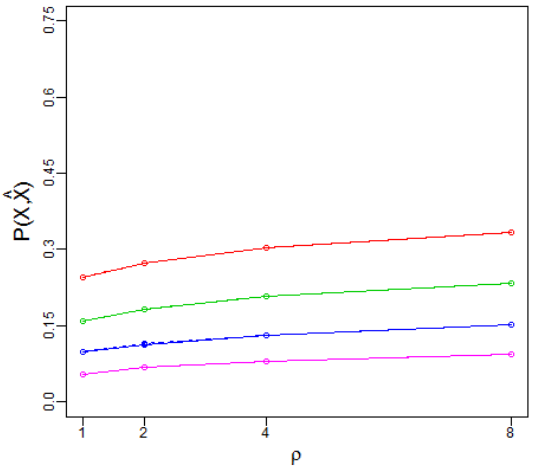



Figure 5.3: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the MBA simulations for a semi-circle. Column one: perturbed distances $\tilde{\mathbf{D}}$ are generated using the exponential transform (4.4). Column two: $\tilde{\mathbf{D}}$ are generated using the power transform (4.6). Row one: $P(\mathbf{X}, \hat{\mathbf{X}})$ values from fitting $\tilde{\mathbf{D}}$ into two dimensional Euclidean space with metric MDS. Row two: $P(\mathbf{X}, \hat{\mathbf{X}})$ values from fitting $\tilde{\mathbf{D}}$ into two dimensional Euclidean space with non-metric MDS. The red lines for $\alpha = 0.1$ (4.4) or $b_0 = 0.1$ (4.6); the green lines for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines for $\alpha = 0.0001$ or $b_0 = 0.0001$. In the power transform figures the solid lines signify the $m_{\text{min}} = 1$ adjustment has been applied, and the dashed lines signify the $m_{\text{min}} = 2$ adjustment has been applied.

(a) Exponential transform & metric MDS. (b) Power transform & metric MDS.

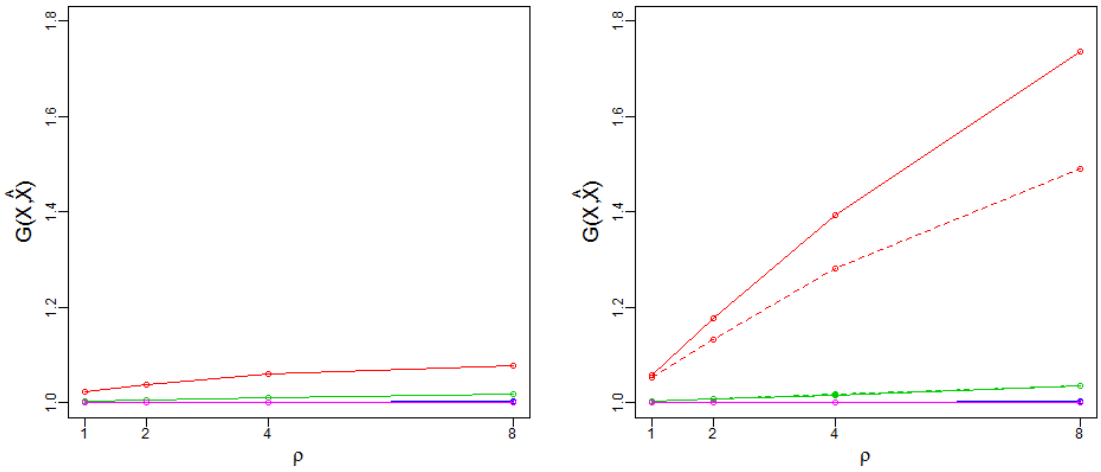



Figure 5.4: Size expansion $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) values from the MBA simulations for a semi-circle, where the perturbed distances $\tilde{\mathbf{D}}$ are fit into two dimensional Euclidean space using metric MDS. Left panel: $\tilde{\mathbf{D}}$ are generated using the exponential transform (4.4). Right panel: $\tilde{\mathbf{D}}$ are generated using the power transform (4.6). The red lines for $\alpha = 0.1$ (4.4) or $b_0 = 0.1$ (4.6); the green lines for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines for $\alpha = 0.0001$ or $b_0 = 0.0001$. In the power transform figures the solid lines signify the $m_{\min} = 1$ adjustment has been applied, and the dashed lines signify the $m_{\min} = 2$ adjustment has been applied.

eigenvalues $\hat{\lambda}_k$ (2.7) and original eigenvalues will be displayed in a scree plot, to provide a similar analysis to the eigenvalue scree used in Section 4.2.1.

**Exponential transform with metric MDS**

The poorest fitting $P(\mathbf{X}, \hat{\mathbf{X}})$ for the exponential transform (4.4) and metric MDS, of those generated occurs at $\alpha = 0.1$ and $\rho = 8$. The $\mathbf{X}$ and a $\hat{\mathbf{X}}$ generated using these parameters are displayed in Figure 5.5, with $\hat{\lambda}_k$ in Figure 5.6. The fitted configuration $\hat{\mathbf{X}}$ in Figure 5.5 appears to have retained its semi-circular structure in the first and second dimensions, with noise forcing the points to meander about the arc. Noise in the third dimension (first spurious dimension) appears more intense with clustering of points at the ends of the configuration.

The principal and secondary fitted eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$ denoted by $\bullet$ in Figure 5.6 appear much larger than spurious non-zero fitted eigenvalues $\hat{\lambda}_k$ for $k = 3, \ldots, 100$ denoted by $\circ$, suggesting retention of structure in these dimensions. This is supported by the magnitude criterion Sibson (1979), which states genuine eigenvalues should have a magnitude greater than the absolute magnitude of the largest negative eigenvalue. The $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are larger than their original eigenvalue counterparts $\lambda_1$ and $\lambda_2$ denoted by the $\bullet$ in Figure 5.6, this difference is most noticeable in the second eigenvalue. The spurious $\hat{\lambda}_k$ appear to decrease in size linearly, with the absolute magnitude of the largest positive and negative eigenvalues almost equal.

**Power transform with metric MDS**

The poorest fitting $P(\mathbf{X}, \hat{\mathbf{X}})$ for the power transform (4.6) and metric MDS, occurs at $b_0 = 0.1$, $\rho = 8$ and $m_{\text{min}} = 1$. The $\mathbf{X}$ and a $\hat{\mathbf{X}}$ generated using these parameters are displayed in Figure 5.7, with $\hat{\lambda}_k$ in Figure 5.8. The fitted configuration $\hat{\mathbf{X}}$ in Figure 5.7 appears to have lost its semi-circular structure but it has retained some linear structure

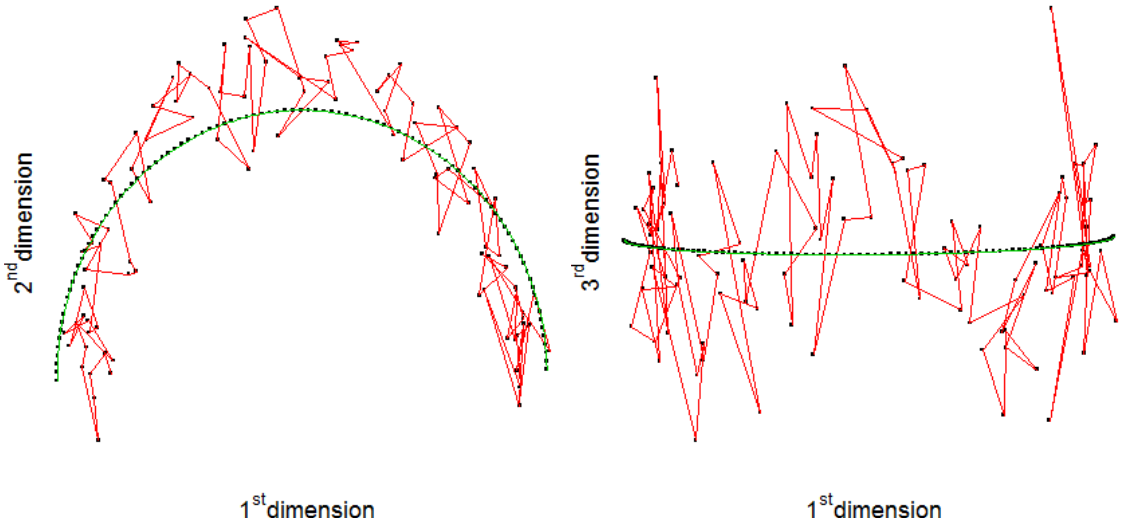



Figure 5.5: Fitted $\hat{\mathbf{X}}$ and original configurations $\mathbf{X}$ for a semi-circle, generated using the MBA approach with the exponential transform (4.4) using $\alpha = 0.1$ and dispersion of $\rho = 8$. Then the $\tilde{\mathbf{D}}$ is fitted into three dimensional Euclidean space using metric MDS. The ●— denotes a point of $\hat{\mathbf{X}}$ and the red line connects successive points of $\hat{\mathbf{X}}$. The ●— denotes a point of $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

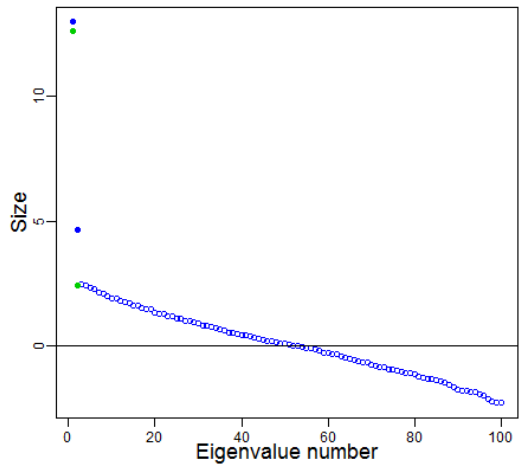



Figure 5.6: Fitted eigenvalues (2.7) from the fitted semi-circle generated in Figure 5.5. The blue circles ● denote the genuine fitted eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$; the hollow blue circles ○ denotes the spurious fitted eigenvalues $\hat{\lambda}_k$ for $k \geq 3$, and the green circles ● denotes the original eigenvalues $\lambda_k$ for $k = 1, 2$.

in the first dimension. The intensity of the noise decreases towards the centre of the configuration, also the intensity of the noise in the third dimension appears to decrease towards the centre of the configuration. This could be due to small $\mu_{i,j}$ shared between points at opposite ends of the configuration being more susceptible to greater noise. The $\hat{\lambda}_1$ in Figure 5.8 appears clear of the spurious $\hat{\lambda}_k$ although $\hat{\lambda}_2$ appears to group with the spurious $\hat{\lambda}_k$. The absolute magnitude of the largest negative $\hat{\lambda}_k$ is approximately equal to the magnitude of $\hat{\lambda}_2$, the magnitude criterion would classify $\hat{\lambda}_2$ as a spurious eigenvalue. This suggests little structure is retained in the second dimension. The $\hat{\lambda}_1$ is much larger than its original counterpart $\lambda_1$, driven by the introduction of noise in $\hat{\lambda}_1$. The spurious eigenvalues decline in an "s" shape.

**Exponential transform with non-metric MDS**

The poorest fitting $P(\mathbf{X}, \hat{\mathbf{X}})$ for the exponential transform (4.4) and non-metric MDS, occurs at $\alpha = 0.1$ and $\rho = 8$. The $\mathbf{X}$ and one $\hat{\mathbf{X}}$ generated using these parameters are displayed in Figure 5.9. The $\hat{\mathbf{X}}$ in Figure 5.9 appears to have retained its semi-circular structure, although the ends appear to have become involuted. Noise appears to distributed evenly, with the points meandering about the outline of the semi-circle. The stress value is $S_2(\hat{\mathbf{X}}) = 29.3032\%$ suggesting a poor fit.

**Power transform with non-metric MDS**

The poorest fitting $P(\mathbf{X}, \hat{\mathbf{X}})$ for the power transform (4.6) and non-metric MDS, occurs at $b_0 = 0.1$; $\rho = 8$ and $m_{\text{min}} = 1$. The $\mathbf{X}$ and one $\hat{\mathbf{X}}$ generated using these parameters are displayed in Figure 5.10. The $\hat{\mathbf{X}}$ in Figure 5.10 has warped into a horseshoe with strong involution of the end points. This horseshoeing could be due to medium distances been perturbed into larger distances. Noise in $\hat{\mathbf{X}}$ appears weaker, with only minor meandering of points in $\hat{\mathbf{X}}$. The stress value $S_2(\hat{\mathbf{X}}) = 11.0878\%$ suggesting a fair fit.

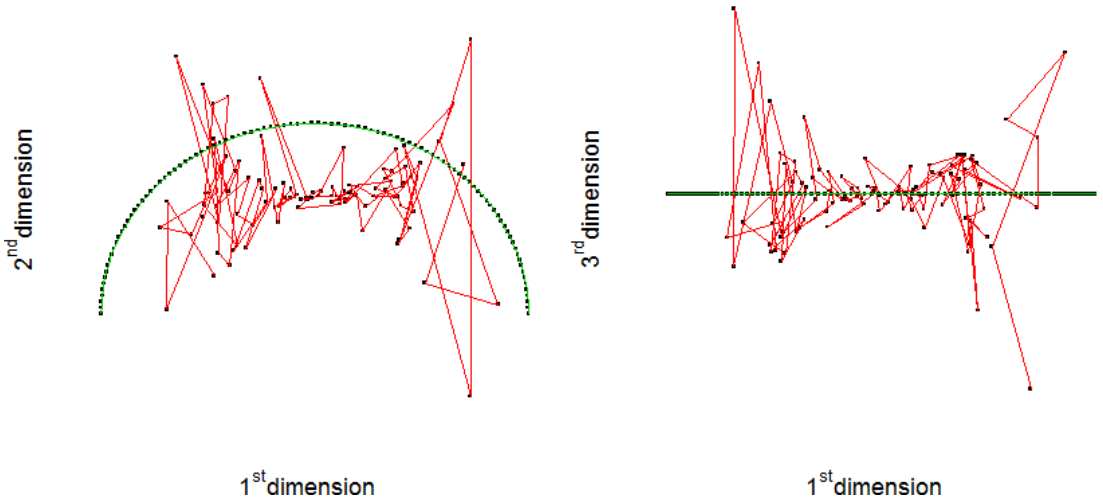



Figure 5.7: Fitted $\hat{\mathbf{X}}$ and original configurations $\mathbf{X}$ for a semi-circle, generated using the MBA approach with the power transform (4.6) using $b_0 = 0.1$ and $\beta = -0.5$; dispersion of $\rho = 8$ and applying the $m_{\min} = 1$ adjustment. Then the $\tilde{\mathbf{D}}$ is fitted into three dimensional Euclidean space using metric MDS. The $\bullet$– denotes a point of $\hat{\mathbf{X}}$ with the red line connecting successive points of $\hat{\mathbf{X}}$. The $\bullet$– denotes a point of $\mathbf{X}$ with the green line connecting successive points of $\mathbf{X}$.

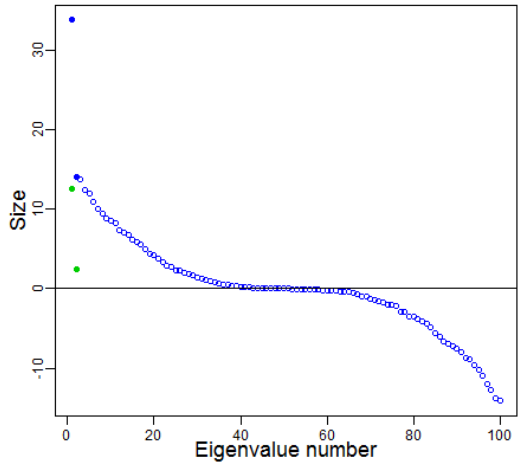



Figure 5.8: Fitted eigenvalues (2.7) from the fitted semi-circle generated in Figure 5.7. The blue circles ● denote the genuine fitted eigenvalues $\hat{\lambda}_1$ and $\hat{\lambda}_2$; the hollow blue circles ○ denotes the spurious fitted eigenvalues $\hat{\lambda}_k$ for $k \geq 3$, and the green circles ● denotes the original eigenvalues $\lambda_k$ for $k = 1, 2$.

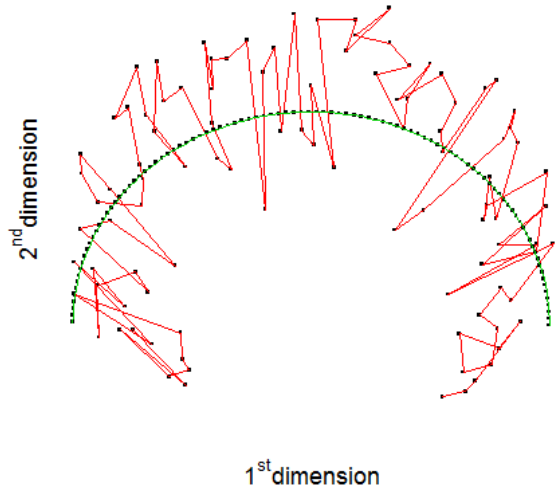



Figure 5.9: Fitted $\hat{\mathbf{X}}$ and original configurations $\mathbf{X}$ for a semi-circle, generated using the MBA approach with the exponential transform (4.4) using $\alpha = 0.1$ and dispersion of $\rho = 8$. Then the $\tilde{\mathbf{D}}$ is fitted into three dimensional Euclidean space using non-metric MDS. The —● denotes a point of $\hat{\mathbf{X}}$ and the red line connects successive points of $\hat{\mathbf{X}}$. The —● denotes a point of $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

## 5.3 Properties of the perturbed distances

The simulation results from the MBA provide insight into how altering parameters affects the fitted configuration $\hat{\mathbf{X}}$. One result is the increase in $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) which is particularly large when using the power transform. The part of the MBA which is investigated here is the structure of the perturbed distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$.

### 5.3.1 Delta method

The $\tilde{d}_{i,j}$ are products of random variables $m_{i,j}$ passed through a function $\tilde{d}_{i,j} = f(m_{i,j})$. Thus the $\tilde{d}_{i,j}$ are also random variables. The Delta method found in Stuart and Ord (1994) pages 350-351 and Rao (1966) pages 319 - 320 can be used to make inferences on the properties of the $\tilde{d}_{i,j}$. The Delta method takes the Taylor-series expansion of the

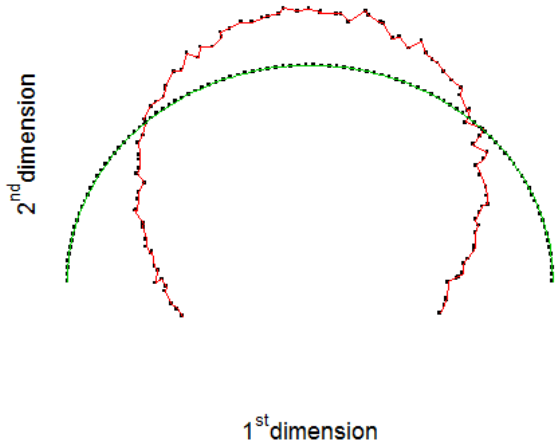



Figure 5.10: Fitted $\hat{\mathbf{X}}$ and original configurations $\mathbf{X}$ for a semi-circle, generated using the MBA approach with the power transform (4.6) using $b_0 = 0.1$ and $\beta = -0.5$; dispersion of $\rho = 8$ and applying the $m_{\text{min}} = 1$ adjustment. Then the $\tilde{\mathbf{D}}$ is fitted into three dimensional Euclidean space using non-metric MDS. The —●— denotes a point of $\hat{\mathbf{X}}$ with the red line connecting successive points of $\hat{\mathbf{X}}$. The —●— denotes a point of $\mathbf{X}$ with the green line connecting successive points of $\mathbf{X}$.

function around the mean of the random variable passed through it. Then the expectation and variance of the expansion are found to give approximate values for $E(f(m_{i,j}))$ and var$(f(m_{i,j}))$.

Taking the Taylor-series expansion of $f(m_{i,j})$ around $\mu_{i,j}$ to second order gives

$$f(m_{i,j}) \approx f(\mu_{i,j}) + f'(\mu_{i,j})(m_{i,j} - \mu_{i,j}) + \frac{f''(\mu_{i,j})}{2!}(m_{i,j} - \mu_{i,j})^2. \tag{5.7}$$

Taking the expectation of (5.7) gives

$$\begin{aligned} E(f(m_{i,j})) &\approx f(\mu_{i,j}) + f'(\mu_{i,j})E(m_{i,j} - \mu_{i,j}) + \frac{f''(\mu_{i,j})}{2!}E((m_{i,j} - \mu_{i,j})^2) \\ &\approx f(\mu_{i,j}) + \frac{f''(\mu_{i,j})}{2!}\text{var}(m_{i,j}) \\ &\approx f(\mu_{i,j}) + \frac{f''(\mu_{i,j})}{2!}\rho\mu_{i,j}. \end{aligned} \tag{5.8}$$

Taking the variance of the first two terms of (5.7) gives

$$\text{var}(f(m_{i,j})) \approx f'(\mu_{i,j})^2\text{var}(m_{i,j}). \tag{5.9}$$

Subtracting the original distances from $E(f(m_{i,j}))$ in (5.8) gives the bias in $\tilde{d}_{i,j}$

$$\begin{aligned} \text{Bias}(\tilde{d}_{i,j}) &= E(f(m_{i,j})) - f(\mu_{i,j}) \\ &= \frac{f''(\mu_{i,j})}{2}\rho\mu_{i,j}. \end{aligned} \tag{5.10}$$

This bias could be responsible for the observed increase in $G(\mathbf{X}, \hat{\mathbf{X}})$, applying the delta method to the exponential transform (4.4) and the power transform (4.6) we can see this clearer.

### 5.3.2 Exponential transform

The Taylor-series expansion of the exponential transform ($f(m_{i,j}) = e^{-\alpha m_{i,j}}$) to second order is

$$f(m_{i,j}) \approx e^{-\alpha\mu_{i,j}} - \alpha e^{-\alpha\mu_{i,j}}(m_{i,j} - \mu_{i,j}) + \frac{\alpha^2}{2}e^{-\alpha\mu_{i,j}}(m_{i,j} - \mu_{i,j})^2. \tag{5.11}$$

Taking the expectation of (5.11) gives

$$\begin{aligned} E(f(m_{i,j})) &\approx e^{-\alpha\mu_{i,j}} + \frac{\alpha^2}{2} e^{-\alpha\mu_{i,j}} \rho\mu_{i,j} \\ &\approx d_{i,j}\left(1 - \frac{\alpha}{2}\log(d_{i,j})\rho\right). \end{aligned} \tag{5.12}$$

Taking the variance of the first two terms of (5.11) gives

$$\text{var}(f(m_{i,j})) \approx \alpha^2 e^{-2\alpha\mu_{i,j}} \rho\mu_{i,j} = -\alpha\rho\log(d_{i,j}) d_{i,j}^2. \tag{5.13}$$

Using (5.12) the bias in the exponential transform is calculated as

$$\text{Bias}(f(m_{i,j})) \approx -\frac{\alpha}{2} d_{i,j}\log(d_{i,j})\rho. \tag{5.14}$$

In addition to the additive bias a proportional (inflation) bias which expands the distances, can be taken from (5.12)

$$\text{Inflation}(f(m_{i,j})) \approx 1 - \frac{\alpha}{2}\log(d_{i,j})\rho. \tag{5.15}$$

where $\text{Inflation}(f(m_{i,j})) > 1$.

Figures 5.11 and 5.12 gives plots of $E(\tilde{d}_{i,j})$; $\text{var}(\tilde{d}_{i,j})$; $\text{Bias}(\tilde{d}_{i,j})$ and $\text{Inflation}(\tilde{d}_{i,j})$ all against $d_{i,j}$. The plots of $\text{var}(\tilde{d}_{i,j})$ and $\text{Bias}(\tilde{d}_{i,j})$ against $d_{i,j}$, display a maximum which occurs at $d_{i,j} = e^{-\frac{1}{2}}$ for $\text{var}(\tilde{d}_{i,j})$ and at $d_{i,j} = e^{-1}$ for $\text{Bias}(\tilde{d}_{i,j})$. These maximums indicates the medium distances are becoming less accurate than the larger distances, and is counter to the intuition that accuracy should decay as distance increases. This maximum could be contributing to the horseshoe effect seen in $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{E,NM}$ in Section 4.2.

(a) $E(\tilde{d}_{i,j})$ vs $d_{i,j}$. (b) var$(\tilde{d}_{i,j})$ vs $d_{i,j}$.

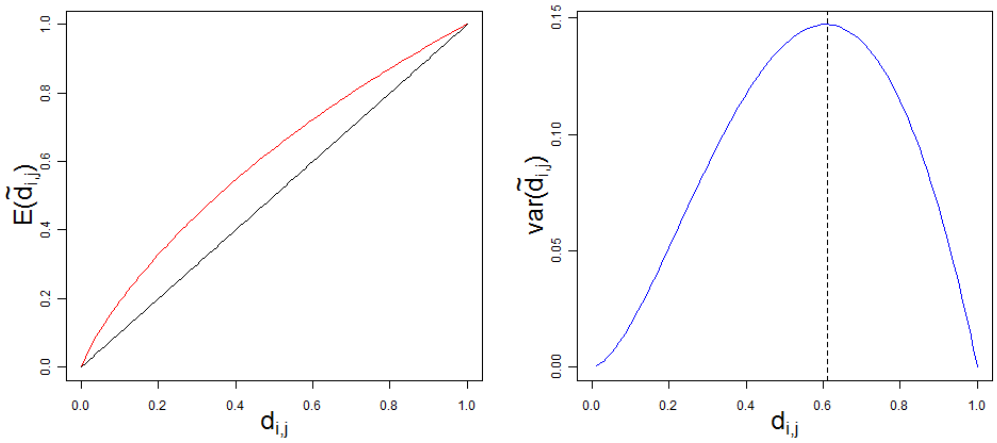



Figure 5.11: Left panel: plot of expected distance for the exponential transform $E(\tilde{d}_{i,j})$ (5.12) at $\alpha = 0.1$ and $\rho = 8$ against original distance $d_{i,j}$, denoted by ——; with the identity line for comparison ——. Right panel: plot of the variance for the exponential transform var$(\tilde{d}_{i,j})$ (5.13) at $\alpha = 0.1$ and $\rho = 8$ against $d_{i,j}$ ——; the dashed line - - - indicated the location of the maximum in the variance at $d_{i,j} = e^{-\frac{1}{2}}$.

(a) Bias$(\tilde{d}_{i,j})$ vs $d_{i,j}$. (b) Inflation$(\tilde{d}_{i,j})$ vs $d_{i,j}$.

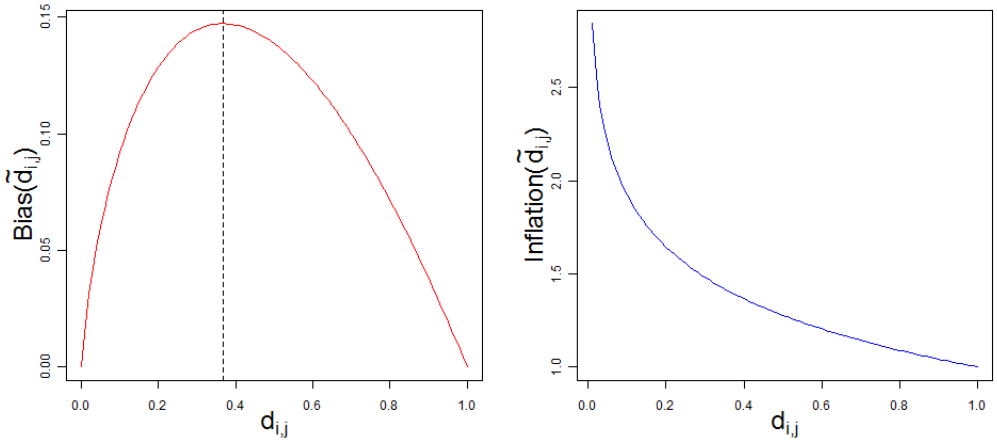



Figure 5.12: Left panel: plot of bias in the distance for the exponential transform Bias$(\tilde{d}_{i,j})$ (5.14) at $\alpha = 0.1$ and $\rho = 8$ against original distance $d_{i,j}$, denoted by ——; the dashed line - - - indicated the location of the maximum in the bias at $d_{i,j} = e^{-1}$. Right panel: plot of inflation in the distances for the exponential transform Inflation$(\tilde{d}_{i,j})$ (5.15) at $\alpha = 0.1$ and $\rho = 8$ against $d_{i,j}$ ——.

### 5.3.3 Power transform

The Taylor-series expansion of the power transform ($f(m_{i,j}) = (b_0 m_{i,j})^{\beta}$) to second order is

$$f(m_{i,j}) \approx b_0^{\beta}\mu_{i,j}^{\beta} + \beta b_0^{\beta}\mu_{i,j}^{\beta-1}(m_{i,j} - \mu_{i,j}) + \frac{\beta}{2}(\beta-1)b_0^{\beta}\mu_{i,j}^{\beta-2}(m_{i,j} - \mu_{i,j})^2. \tag{5.16}$$

The effect of the $m_{\text{min}}$ adjustment on $E(m_{i,j})$ and var$(m_{i,j})$ is ignored for simplicity. Taking the expectation of (5.16) gives

$$\begin{aligned} E(f(m_{i,j})) &\approx b_0^{\beta}\mu_{i,j}^{\beta} + \frac{\beta}{2}(\beta-1)b_0^{\beta}\rho\mu_{i,j}^{\beta-1} \\ &\approx d_{i,j}\left(1 + \frac{\beta}{2}(\beta-1)\rho b_0 d_{i,j}^{-\frac{1}{\beta}}\right). \end{aligned} \tag{5.17}$$

Taking the variance of the first two terms of (5.16) gives

$$\text{var}(f(m_{i,j})) \approx \beta^2 b_0^{2\beta}\rho\mu_{i,j}^{2\beta-1} = d_{i,j}^{2-\frac{1}{\beta}} b_0 \beta^2 \rho. \tag{5.18}$$

Using (5.17) the bias in the power transform is calculated as

$$\text{Bias}(f(m_{i,j})) \approx \frac{\beta}{2}(\beta-1)\rho b_0 d_{i,j}^{1-\frac{1}{\beta}}. \tag{5.19}$$

and a proportional (inflationary) bias is calculated as

$$\text{Inflation}(f(m_{i,j})) \approx 1 + \frac{\beta}{2}(\beta-1)\rho\mu_{i,j}^{-1}. \tag{5.20}$$

where Inflation$(f(m_{i,j})) > 1$.

The plots of var$(f(m_{i,j}))$ and Bias$(f(m_{i,j}))$ against $d_{i,j}$ in Figure 5.13 and 5.14 reveal

a maximum at the boundary of $d_{i,j}$ used in the original configurations ($d_{i,j} = 1$). Differentiation shows the maximum occurs at $d_{i,j} = \infty$, hence accuracy continues to decrease as distance increases. The decrease in accuracy as distance increases is what is expected in a sensible transform function.

## 5.4 Unbiased simulations

To gauge the effect of bias in the perturbed distances from (5.14) and (5.19) on the fitted configuration $\hat{\mathbf{X}}$, the simulations of the MBA were repeated using unbiased perturbed distances $\tilde{\tilde{\mathbf{D}}} = (\tilde{\tilde{d}}_{i,j})$. The $\tilde{\tilde{d}}_{i,j}$ were such that $E(\tilde{\tilde{d}}_{i,j}) = d_{i,j}$ and var$(\tilde{\tilde{d}}_{i,j}) =$ var$(\tilde{d}_{i,j})$, and were generated using the normal distribution with mean $d_{i,j}$ and variance var$(\tilde{d}_{i,j})$. Studying the difference between the MBA (biased) and unbiased simulation results should provide insight into the effect of the bias on $\hat{\mathbf{X}}$, and studying the unbiased simulation results gave insight into the effect of the variance on $\hat{\mathbf{X}}$.

### Exponential transform

Unbiased perturbed distances emulating the exponential transform were simulated using

$$\tilde{\tilde{d}}_{i,j} \sim N(d_{i,j}, -\alpha\rho \log(d_{i,j}) d_{i,j}^2), \tag{5.21}$$

where the variance parameter is from (5.13), and any $\tilde{\tilde{d}}_{i,j} \leq 0$ were replaced with $d_{i,j}$ and all $\tilde{\tilde{d}}_{i,j} > 1$ were unaltered.

(a) $E(\tilde{d}_{i,j})$ vs $d_{i,j}$. (b) var$(\tilde{d}_{i,j})$ vs $d_{i,j}$.

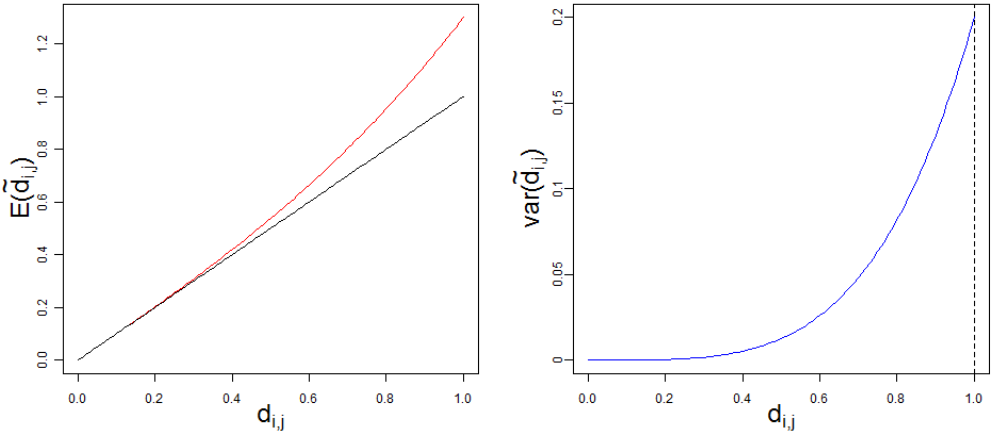



Figure 5.13: Left panel: plot of expected distance for the power transform $E(\tilde{d}_{i,j})$ (5.17) at $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$ against original distance $d_{i,j}$, denoted by ——; with the identity line for comparison ——. Right panel: plot of the variance for the power transform var$(\tilde{d}_{i,j})$ (5.18) at $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$ against $d_{i,j}$ ——; the dashed line - - - indicated the location of the maximum in the variance at $d_{i,j} = 1$.

(a) Bias$(\tilde{d}_{i,j})$ vs $d_{i,j}$ (b) Inflation$(\tilde{d}_{i,j})$ vs $d_{i,j}$

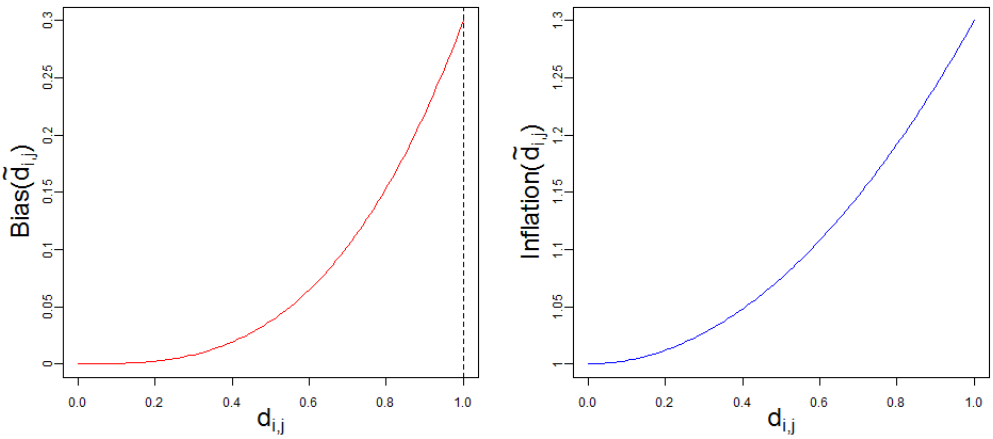



Figure 5.14: Left panel: plot of bias in the distance for the power transform Bias$(\tilde{d}_{i,j})$ (5.19) at $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$ against original distance $d_{i,j}$, denoted by ——; the dashed line - - - indicated the location of the maximum in the bias at $d_{i,j} = 1$. Right panel: plot of inflation in the distances for the power transform Inflation$(\tilde{d}_{i,j})$ (5.20) at $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$ against $d_{i,j}$ ——.

### Power transform

Unbiased perturbed distances emulating the power transform were simulated using

$$\tilde{\tilde{d}}_{i,j} \sim N(d_{i,j}, d_{i,j}^{2-\frac{1}{\beta}} b_0 \beta^2 \rho), \tag{5.22}$$

where the variance parameter is from (5.18), and any $\tilde{\tilde{d}}_{i,j} \leq 0$ were replaced with $d_{i,j}$. The unbiased simulations for both transforms were run in a identical manor to the MBA simulations, but without the $m_{\text{min}}$ adjustment as no counts were used.

## 5.4.1 Unbiased simulation results

The unbiased simulation results are presented in a similar format to the MBA simulation results, with the semi-circle results plotted, and comments made incorporating simulations from all the shapes. The unbiased simulation results can be found in Appendix Section C.

### MDS performance

Comparing the metric biased and unbiased $\theta_{1:p}$ (2.12) simulation results, the unbiased simulations show slightly worst performance for the exponential transform compared to the power transform. Comparing the non-metric biased and unbiased $S_p(\hat{\mathbf{X}})$ (2.14) simulation results, the unbiased simulations show worse performance of the exponential transform compared to the power transform. The margin between the biased and unbiased MDS performance narrows as $C_v(\mu_{i,j}, \rho)$ (5.4) decreases. As the size of the bias decreases as $C_v(\mu_{i,j}, \rho)$ decreases. Studying the unbiased simulations the variance appears to be a more dominant feature than the bias in influencing MDS performance.

## Shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$

The shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) could be a useful measure for correcting bias, as it is applicable to $\hat{\mathbf{X}}$ from both metric and non-metric MDS. Comparing the metric biased and unbiased $P(\mathbf{X}, \hat{\mathbf{X}})$ simulation results, the unbiased simulations show poorer performance for the exponential transform but better performance for the power transform. The margin between biased and unbiased $P(\mathbf{X}, \hat{\mathbf{X}})$ for the power transform and metric MDS is quite wide, indicating its potential for techniques of bias correction. Comparing the non-metric biased and unbiased $P(\mathbf{X}, \hat{\mathbf{X}})$ simulation results, removing the bias makes little difference to the exponential transforms results and only improves the power transform results when $C_v(\mu_{i,j}, \rho)$ (5.4) is large, suggesting non-metric MDS is quite robust to bias. As $C_v(\mu_{i,j}, \rho)$ decreases the margin between biased and unbiased $P(\mathbf{X}, \hat{\mathbf{X}})$ diminishes. Studying the unbiased simulation results the variance appears to be a more dominant feature than the bias in increasing $P(\mathbf{X}, \hat{\mathbf{X}})$.

## Size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$

Comparing the (metric) biased and unbiased $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) simulation results, for the exponential transform the removal of bias makes little difference to $G(\mathbf{X}, \hat{\mathbf{X}})$, but for the power transform, removing bias significantly improves $G(\mathbf{X}, \hat{\mathbf{X}})$. The improvement with the power transform indicates it is a candidate for a bias correction. The margin between biased and unbiased $G(\mathbf{X}, \hat{\mathbf{X}})$ decreases as $C_v(\mu_{i,j}, \rho)$ decreases. Studying the unbiased simulation results the variance appears to be a more dominant feature than the bias in increasing $G(\mathbf{X}, \hat{\mathbf{X}})$.

### Unbiased simulations summary

Of the four routes (using either the exponential transform (4.4) or power transform (4.6), and either metric or non-metric MDS) to obtain an estimated chromosome configuration $\hat{\mathbf{X}}$ in Chapter 4, the power transform with metric MDS, appears to be in strongest need for bias correction. The exponential transform with metric or non-metric MDS and the power transform with non-metric MDS do not appear so influenced by bias. The variance is a major component in causing a poor $\hat{\mathbf{X}}$: an approach to reduce this variance effect is discussed in Section 6.4.

## 5.4.2 Validlity of unbiased simulation

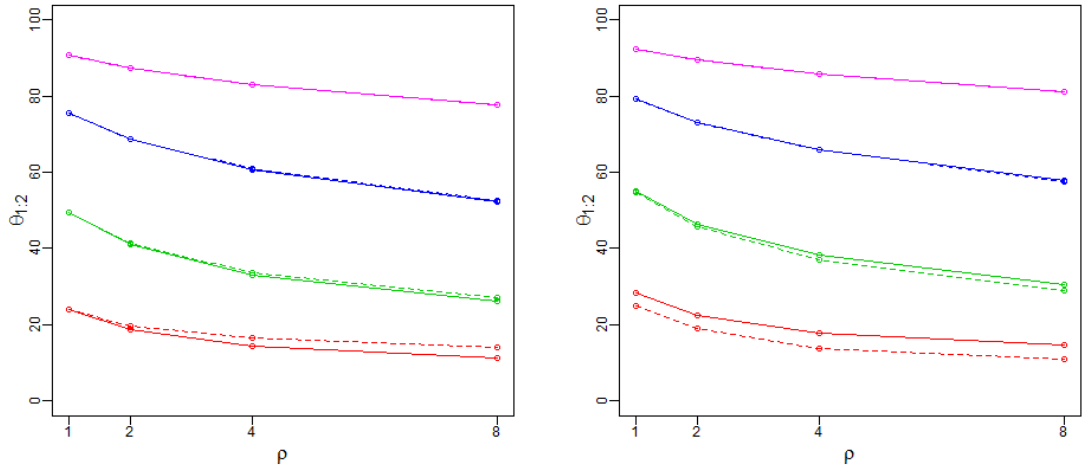
The validity of the unbiased simulations depends on how well $\tilde{\tilde{\mathbf{D}}}$ emulates $\tilde{\mathbf{D}}$, without the presence of a bias. To assess validity, Shepards plots were used with $\tilde{\tilde{d}}_{i,j}$ generated when $C_v(\mu_{i,j}, \rho)$ is either large or moderate. The plots included a set of $\tilde{d}_{i,j}$ generated under the same levels of $C_v(\mu_{i,j}, \rho)$. The addition of an identity line, and a 95% confidence interval for the $\tilde{\tilde{d}}_{i,j}$ around the identity line to the Shepards plots, should aid interpretation. The array of information in the Shepards plots indicates where $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$ match and where they differ, how $\tilde{\tilde{d}}_{i,j}$ is distributed and whether there are any other tell-tale signs of poor emulation. In addition to the Shepards plots, the percentage of $\tilde{\tilde{d}}_{i,j}$ simulated below zero and above one is calculated for the exponential transform (4.4) and percentage of the $\tilde{\tilde{d}}_{i,j}$ simulated below zero is calculated for the power transform (4.6). The Shepards plots and proportions use data from a semi-circle configuration and a single simulation.

### Exponential transform

The conditions which lead to large $C_v(\mu_{i,j}, \rho)$ for the exponential transform (4.4), are $\alpha = 0.1$ and $\rho = 8$, and conditions which lead to moderate $C_v(\mu_{i,j}, \rho)$ are $\alpha = 0.01$ and

(a) Exponential transform & metric MDS.

(b) Power transform & metric MDS.




(c) Exponential transform & non-metric MDS.

(d) Power transform & non-metric MDS.

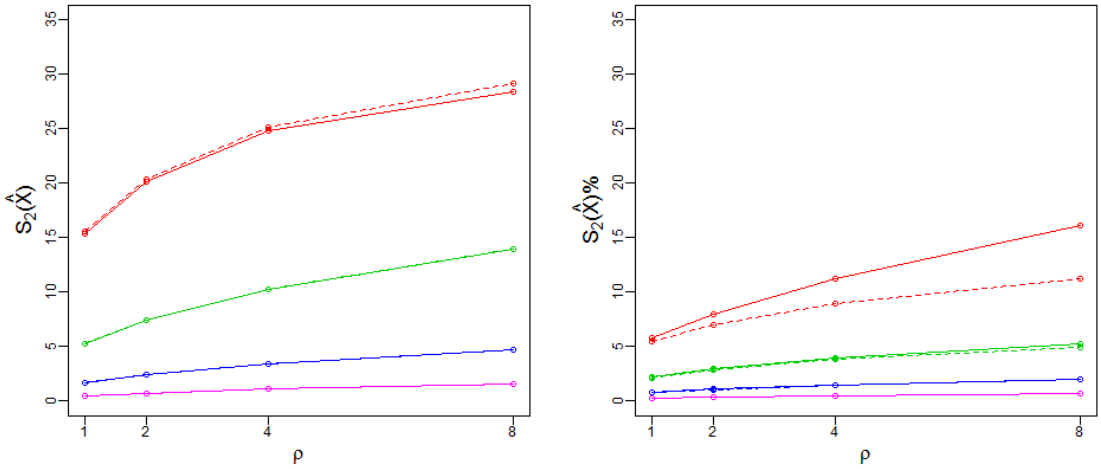
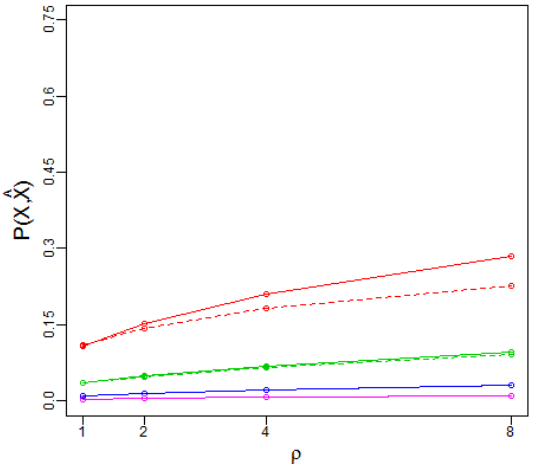
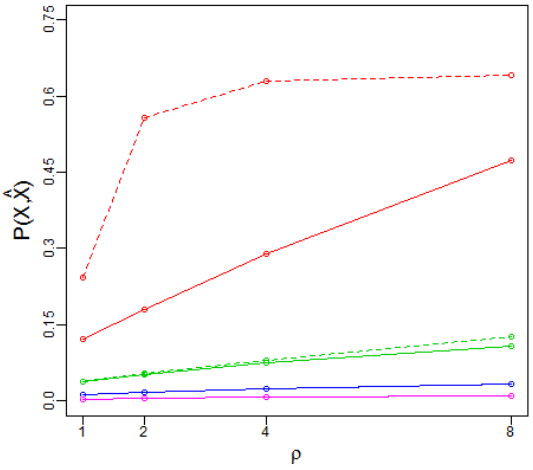



Figure 5.15: MDS performance statistics $\theta_{1:2}$ (2.12) or $S_2(\hat{\mathbf{X}})$ (2.14) from the unbiased MBA simulations for a semi-circle. Column one: unbiased perturbed distances $\tilde{\tilde{\mathbf{D}}}$ (5.21) emulating the exponential transform (4.4). Column two: $\tilde{\tilde{\mathbf{D}}}$ (5.22) emulating the power transform (4.6). Row one: $\theta_{1:2}$ values from fitting $\tilde{\tilde{\mathbf{D}}}$ into two dimensional Euclidean space with metric MDS. Row two: $S_2(\hat{\mathbf{X}})$ values from fitting $\tilde{\tilde{\mathbf{D}}}$ into two dimensional Euclidean space with non-metric MDS. The red lines —⊖— for $\alpha = 0.1$ (5.21) or $b_0 = 0.1$ (5.22); the green lines —⊖— for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines —⊖— for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines —⊖— for $\alpha = 0.0001$ or $b_0 = 0.0001$. The dashed lines - ⊖ - give the equivalent MBA simulation values from Figure 5.2.

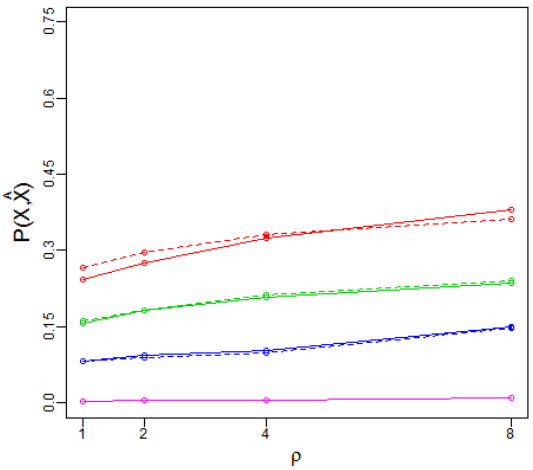
(a) Exponential transform & metric MDS.




(b) Power transform & metric MDS.




(c) Exponential transform & non-metric MDS.




(d) Power transform & non-metric MDS.

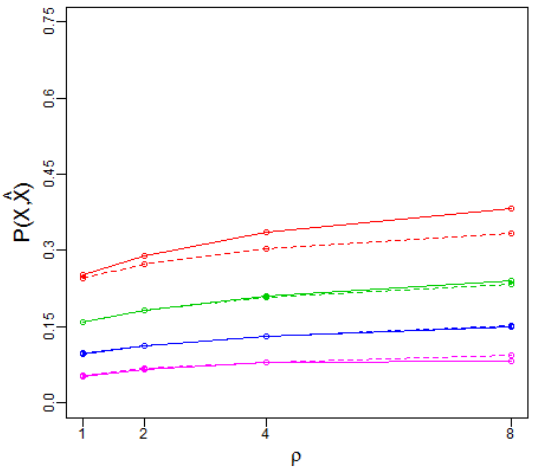



Figure 5.16: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the unbiased MBA simulations for a semi-circle. Column one: unbiased perturbed distances $\tilde{\tilde{\mathbf{D}}}$ (5.21) emulating the exponential transform (4.4). Column two: $\tilde{\tilde{\mathbf{D}}}$ (5.22) emulating the power transform (4.6). Row one: $P(\mathbf{X}, \hat{\mathbf{X}})$ values from fitting $\tilde{\tilde{\mathbf{D}}}$ into two dimensional Euclidean space with metric MDS. Row two: $P(\mathbf{X}, \hat{\mathbf{X}})$ values from fitting $\tilde{\tilde{\mathbf{D}}}$ into two dimensional Euclidean space with non-metric MDS. The red lines —○— for $\alpha = 0.1$ (5.21) or $b_0 = 0.1$ (5.22); the green lines —○— for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines —○— for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines —○— for $\alpha = 0.0001$ or $b_0 = 0.0001$. The dashed lines -○- give the equivalent MBA simulation values from Figure 5.3.

(a) Exponential transform & metric MDS. (b) Power transform & metric MDS.

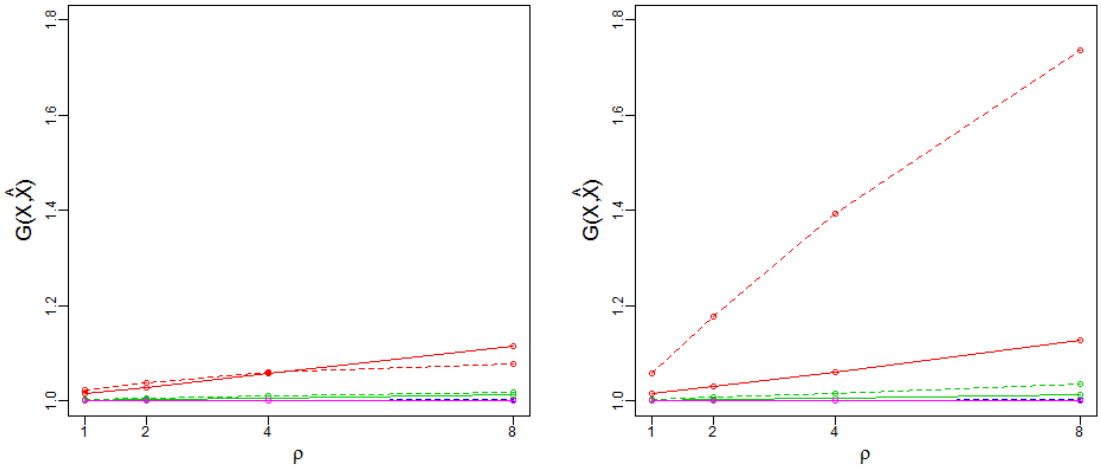



Figure 5.17: Size expansion $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) values from the unbiased MBA simulations for a semi-circle, where the unbiased perturbed distances $\tilde{\tilde{\mathbf{D}}}$ are fit into two dimensional Euclidean space using metric MDS. Left panel: unbiased perturbed distances $\tilde{\tilde{\mathbf{D}}}$ (5.21) emulating the exponential transform (4.4). Right panel: $\tilde{\tilde{\mathbf{D}}}$ (5.22) emulating the power transform (4.6). The red lines —⊖— for $\alpha = 0.1$ (5.21) or $b_0 = 0.1$ (5.22); the green lines —⊖— for $\alpha = 0.01$ or $b_0 = 0.01$; the blue lines —⊖— for $\alpha = 0.001$ or $b_0 = 0.001$, and the pink lines —⊖— for $\alpha = 0.0001$ or $b_0 = 0.0001$. The dashed lines -⊖- give the equivalent MBA simulation values from Figure 5.4.

$\rho = 4$.

(a) Large $C_v(\mu_{i,j}, \rho)$. (b) Moderate $C_v(\mu_{i,j}, \rho)$.

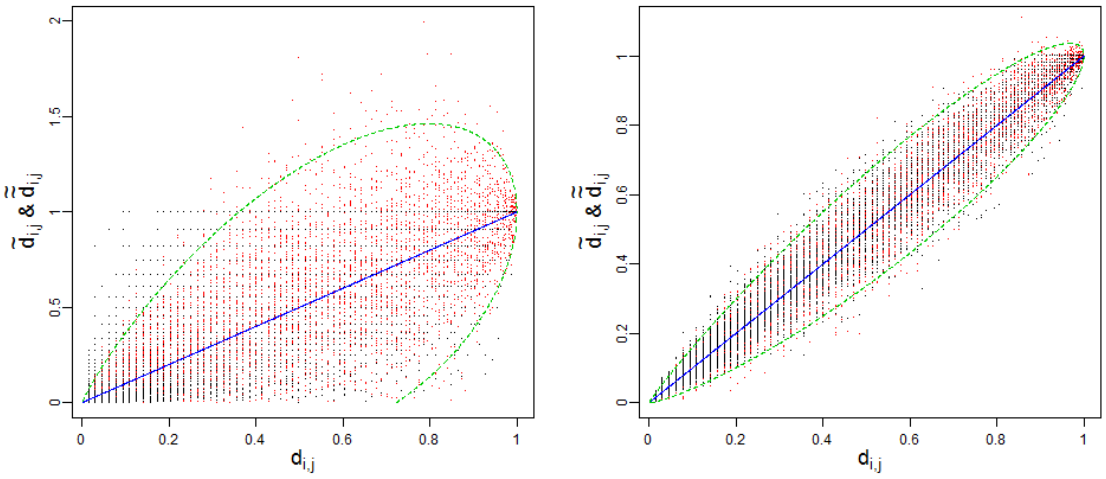



Figure 5.18: Shepards plots for the exponential transform (4.4) MBA perturbed distances $\tilde{\mathbf{D}}$ and unbiased distances $\tilde{\tilde{\mathbf{D}}}$ (5.21) emulating the exponential transform. Left panel: both sets of distances generated using $\alpha = 0.1$ and $\rho = 8$; when coefficient of variation $C_v(\mu_{i,j}, \rho)$ (5.4) is large. Right panel: both sets of distances generated using $\alpha = 0.01$ and $\rho = 4$; when $C_v(\mu_{i,j}, \rho)$ is moderate. The red points . denote elements of $\tilde{\tilde{\mathbf{D}}}$, and the black points . denote elements of $\tilde{\mathbf{D}}$. The blue line — is the identity line and dashed green line - - - is the 95% confidence interval for $\tilde{\tilde{\mathbf{D}}}$ found using var$(\tilde{d}_{i,j})$ (5.13).

| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
|---|---|---|---|---|
| 0.1 | 3.8251 | 7.6817 | 14.2124 | 23.2878 |
| 0.01 | 0.4466 | 0.8420 | 1.5679 | 3.0533 |
| 0.001 | 0.0584 | 0.1045 | 0.1976 | 0.3653 |
| 0.0001 | 0.0061 | 0.0126 | 0.0258 | 0.0464 |

Table 5.2: Percentage of unbiased distances $\tilde{\tilde{d}}_{i,j}$ (5.21) simulated outside the boundary for the exponential transform (4.4) of $\tilde{\tilde{d}}_{i,j} < 0$ and $\tilde{\tilde{d}}_{i,j} > 1$, at different levels of $\alpha$ and dispersion $\rho$.

The Shepards plots for large $C_v(\mu_{i,j}, \rho)$ in Figure 5.18a is a good example of poor emulation. Here the $\tilde{\tilde{d}}_{i,j}$ do not match the $\tilde{d}_{i,j}$ at two locations on the plot. The first

location a large quantity of $\tilde{d}_{i,j}$ lie outside the upper 95% confidence interval of the $\tilde{\tilde{d}}_{i,j}$ around $0 < d_{i,j} \leq 0.3$. The second location is large quantity of $\tilde{\tilde{d}}_{i,j} > 1$. This failure to match is partly due to the absence of a boundary on $\tilde{\tilde{d}}_{i,j} > 1$ and poor variance estimates at large $C_v(\mu_{i,j}, \rho)$. The Shepards plot at medium $C_v(\mu_{i,j}, \rho)$ in Figure 5.18b is an better example of good emulation. Here both $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$ fit inside the 95% confidence interval for $\tilde{\tilde{d}}_{i,j}$, the only location where $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$ fail to match is several $\tilde{d}_{i,j}$ are simulated outside the upper 95% confidence interval, although this is expected due to the influence of the bias. In Table 5.2, when $C_v(\mu_{i,j}, \rho)$ is large at $\alpha = 0.1$ and $\rho \geq 2$, over 5% of the $\tilde{\tilde{d}}_{i,j}$ are simulated outside the boundary, which suggests these simulations should be discounted.

The unusual nature of the exponential transform (4.4) with the bias and variance peaking within the range of the $d_{i,j} = (0, 1]$ and the upper bound on the size of $\tilde{d}_{i,j}$, makes it difficult to produce unbiased perturbed distances which can emulate it. If repeating the unbiased simulations, the lognormal distribution might be more appropriate to simulate $\tilde{\tilde{d}}_{i,j}$ as this removes the possibility simulating $\tilde{\tilde{d}}_{i,j} < 0$.

## Power transform

The conditions which produce large $C_v(\mu_{i,j}, \rho)$ for the power transform (4.6), are $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$ and moderate $C_v(\mu_{i,j}, \rho)$ are $b_0 = 0.01$, $\beta = -0.5$ and $\rho = 4$. The Shepards plots at large $C_v(\mu_{i,j}, \rho)$ in Figure 5.19a, displays a good example of $\tilde{\tilde{\mathbf{D}}}$ emulating $\tilde{\mathbf{D}}$ without the presence of bias. Here there are two locations where $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$ fail to match. The first location is the $\tilde{d}_{i,j}$ simulated above the upper 95% confidence interval for $\tilde{\tilde{d}}_{i,j}$. The second is the absence of $\tilde{d}_{i,j}$ around the lower 95% confidence interval for $\tilde{\tilde{d}}_{i,j}$. These two discrepancies could be due to the presence of the bias driving the $\tilde{d}_{i,j}$ to become larger. At small $d_{i,j}$, both $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$ match very well. The Shepards plot for medium $C_v(\mu_{i,j}, \rho)$ in Figure 5.19b displays good emulation with

(a) Large $C_v(\mu_{i,j}, \rho)$. (b) Moderate $C_v(\mu_{i,j}, \rho)$.

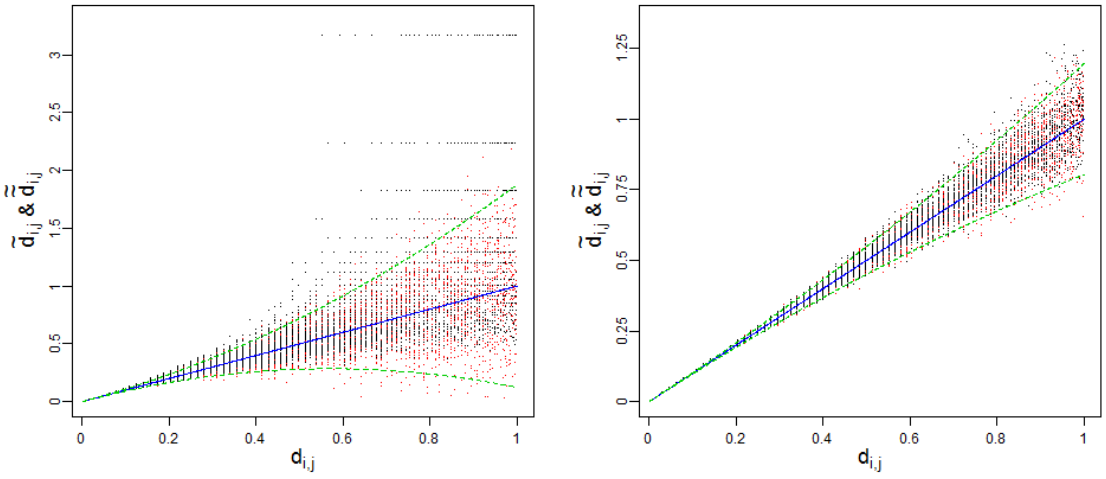



Figure 5.19: Shepards plots for the power transform (4.6) MBA perturbed distances $\tilde{\mathbf{D}}$ and unbiased distances $\tilde{\tilde{\mathbf{D}}}$ (5.22) emulating the power transform. Left panel: both sets of distances generated using $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$; when coefficient of variation $C_v(\mu_{i,j}, \rho)$ (5.4) is large. Right panel: both sets of distances generated using $b_0 = 0.01$, $\beta = -0.5$ and $\rho = 4$; when $C_v(\mu_{i,j}, \rho)$ is moderate. The red points . denote elements of $\tilde{\tilde{\mathbf{D}}}$, and the black points . denote elements of $\tilde{\mathbf{D}}$. The blue line ___ is the identity line and dashed green line - - - is the 95% confidence interval for $\tilde{\tilde{\mathbf{D}}}$ found using var$(\tilde{d}_{i,j})$ (5.18).

| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
|---|---|---|---|---|
| 0.1 | 000000 | 000000 | 0.0053 | 0.1374 |
| 0.01 | 000000 | 000000 | 000000 | 000000 |
| 0.001 | 000000 | 000000 | 000000 | 000000 |
| 0.0001 | 000000 | 000000 | 000000 | 000000 |

Table 5.3: Percentage of unbiased distances $\tilde{\tilde{d}}_{i,j}$ (5.22) emulating the power transform, simulated outside the boundary for the power transform (4.6) of $\tilde{\tilde{d}}_{i,j} < 0$, with $\beta = -0.5$ and at different levels of $b_0$ and dispersion $\rho$.

little difference between $\tilde{\tilde{d}}_{i,j}$ and $\tilde{d}_{i,j}$. In Table 5.3, only at the largest $C_v(\mu_{i,j}, \rho)$, where $b_0 = 0.1, \beta = -0.5$ and $\rho \geq 4$, is a tiny proportion of the $\tilde{\tilde{d}}_{i,j}$ simulated outside the boundary.

The Shepards plots and proportions of $\tilde{\tilde{d}}_{i,j}$ simulated outside the boundary suggest the $\tilde{\tilde{d}}_{i,j}$ emulate $\tilde{d}_{i,j}$ very well and all the unbiased simulations can be counted.

## 5.5 Estimating dispersion

Before a bias correction can be applied to the power transforms (4.6) perturbed distances, an estimate for dispersion $\rho$ is required. The dispersion is a component in the additive bias and inflationary bias of both transforms, so whichever approach is taken to correct the bias a suitable estimate of $\rho$ is required.

Given that $\rho$ is assumed uniform across $\mathbf{M}$, a simple way to estimate $\rho$ when using the power transform and metric MDS would be to use a modification of the $\chi^2$ (4.9) score function

$$\hat{\rho} = \frac{2}{n(n-1)} \sum_{i<j} \frac{(m_{i,j} - \mu_{i,j})^2}{\mu_{i,j}}, \tag{5.23}$$

where $\frac{n(n-1)}{2}$ is the size of the upper triangle of the matrix summed over. Each value contributing to the sum of (5.23) represents an elementwise estimate of $\rho$, which is made robust by taking the mean over the matrix. Taking the expectation of (5.23) gives

$$\begin{aligned} E(\hat{\rho}) &= \frac{2}{n(n-1)} \sum_{i<j} \frac{E((m_{i,j} - \mu_{i,j})^2)}{\mu_{i,j}} \\ &= \frac{2}{n(n-1)} \sum_{i<j} \frac{\rho\mu_{i,j}}{\mu_{i,j}} \\ &= \rho, \end{aligned}$$

hence (5.23) is an unbiased estimator. When using $\mathbf{U}$ and $\mathbf{M}$ in (5.23), good estimates for $\rho$ are obtained. Unfortunately $\mathbf{U}$ is assumed unknown so the fitted counts $\hat{\mathbf{U}} = (\hat{\mu}_{i,j})$ (found using the technique outlined in Chapter 4) were used instead. Using $\hat{\mathbf{U}}$ gave very

large $\hat{\rho}$ as can be seen in Table 5.4, with even $\hat{\mathbf{U}}$ from very small $C_v(\mu_{i,j}, \rho)$ giving poor results. The $\hat{\rho}$ have been inflated by the very large $\hat{\mu}_{i,j}$ caused by decreases in the small distances. To try overcome this inflation, a similar estimate based on (5.18) was used, this also suffered from very large $\hat{\rho}$. Finally an algorithm to find $\hat{\rho}$ using a modification of the stress score function $S_p(\hat{\mathbf{X}})$ (2.14) as a point estimator was developed and is described in Section 5.5.1.

The stress $S_p(\hat{\mathbf{X}})$ isotonically regresses the perturbed (or estimated) distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$ onto the fitted distances $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ (2.2), so they are relative in scale to $\hat{\mathbf{D}}$. When using metric MDS the scale in the fitted distances is preserved so isotonic regression of the estimated distances can be ignored. Avoiding the isotonic regression gives a new "simple stress" score function

$$R_p(\hat{\mathbf{X}}) = \frac{\sum_{i<j}(\tilde{d}_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} \hat{d}_{i,j}^2}. \tag{5.24}$$

where $p$ denotes how many dimensions have been used to find $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ in (2.2).

| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
|---|---|---|---|---|
| 0.1 | 75148.73 | 340634.31 | 645784.82 | 1061448.00 |
| 0.01 | 22154.67 | 38895.33 | 74542.88 | 177882.90 |
| 0.001 | 31561.40 | 69040.25 | 99582.07 | 177674.80 |
| 0.0001 | 25323.80 | 52945.44 | 105143.71 | 229512.80 |

Table 5.4: Estimates for dispersion $\hat{\rho}$ using the fitted counts $\hat{\mathbf{U}}$ from a fitted semi-circle $\hat{\mathbf{X}}$, found using the power transform (4.6) and metric MDS. The $\hat{\rho}$ values are found by inputting the elements of $\hat{\mathbf{U}}$ into (5.23) instead of the mean counts. The process of extracting $\hat{\mathbf{U}}$ from $\hat{\mathbf{X}}$ is outlined in Section 4.1.2.

### 5.5.1 Dispersion estimation algorithm

1. Calculate $R_p(\hat{\mathbf{X}})$ (5.24) for $\hat{\mathbf{X}}$, where $\hat{\mathbf{X}}$ is found using the power transform (4.6) and metric MDS.

2. Choose an value of $\hat{\rho}$ as the initial estimate for the dispersion (a good starting value is $\rho = 1$) and using $\hat{\mathbf{U}}$ as the matrix of mean counts simulate a new matrix of perturbed counts $\tilde{\mathbf{M}} = (\tilde{m}_{i,j})$

$$\tilde{m}_{i,j} \sim \text{NB}(r, l), \text{where } r = \frac{\hat{\mu}_{i,j}}{\hat{\rho} - 1} \text{ and } l = \frac{1}{\hat{\rho}}.$$

If $\hat{\rho} = 1$ use $\tilde{m}_{i,j} \sim \text{Poisson}(\hat{\mu}_{i,j})$. Transform $\tilde{\mathbf{M}}$ into $\tilde{\hat{\mathbf{D}}} = (\tilde{\hat{d}}_{i,j})$ using the power transform (4.6) with the same parameters used to find $\hat{\mathbf{X}}$, and calculate a new simple stress value

$$\tilde{R} = \frac{\sum_{i<j}(\tilde{\hat{d}}_{i,j} - \hat{d}_{i,j})^2}{\sum_{i<j} \hat{d}_{i,j}^2}. \tag{5.25}$$

Repeat $1000$ times to obtain a sample of $\tilde{R}$ (5.25), calculate the sample mean $\overline{\tilde{R}} = 1000^{-1} \sum_{i=1}^{m} \tilde{R}_i$ and then find $\delta_{\hat{\rho}} = |R_p(\hat{\mathbf{X}}) - \overline{\tilde{R}}|$.

3. Repeat step 2. using a different value for $\hat{\rho}$ each time. Eventually choosing the $\hat{\rho}$ which gives the smallest $\delta_{\hat{\rho}}$.

The algorithm to find $\hat{\rho}$ is computationally intensive, to ease computation an interval where $\hat{\rho}$ is expected to lie can be scanned across. The interval should start at $\hat{\rho} = 1$ and end at some large $\hat{\rho}$ such that some structural information is retained in $\tilde{\mathbf{M}}$.

The logic behind the algorithm is that if $\hat{\mathbf{D}} \approx \mathbf{D}$ and $\hat{\rho} \approx \rho$ then $\tilde{\hat{\mathbf{D}}}$ should have similar structural properties to $\tilde{\mathbf{D}}$ and generate a $\tilde{R} \approx R_p(\hat{\mathbf{X}})$.

### 5.5.2 Dispersion estimation algorithm results

The algorithm was trialled using the original distances $\mathbf{D}$ for $\hat{\mathbf{D}}$ in $R_p(\hat{\mathbf{X}})$ (5.24) and $\tilde{\mathbf{M}}$ generated from the mean counts $\mathbf{U}$ to test its performance. The algorithm recovered a

$\hat{\rho} \approx \rho$ on each level of dispersion $\rho = 1, 2, 4,$ and $8$. The algorithm was applied to the actual data generated using the MBA for each level of $m_{\text{min}}$.

The $\hat{\rho}$ estimates in Tables 5.5 and 5.6 are an improvement on those using (5.23). The $\hat{\rho}$ values vary depending on the shape the perturbed counts $\mathbf{M}$ originated from. The parabola appears to give the poorest $\hat{\rho}$, with the $\hat{\rho}$ values from moderate $C_v(\mu_{i,j}, \rho)$ giving poorer estimates than the $\hat{\rho}$ values from large $C_v(\mu_{i,j}, \rho)$. The algorithm scans across an interval, so the poorest dispersion estimates used could be a value from the end of the interval. This would be $\hat{\rho} = 1$ from the lower end of the interval, or a value from the upper end of the interval. If the interval is wisely chosen, the poorest $\hat{\rho}$ would not be so large as to erase structure from $\mathbf{M}$.

| | Straight line | | | | Parabola | | | |
|---|---|---|---|---|---|---|---|---|
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0439 | 1.7473 | 2.8128 | 3.7023 | 1.8583 | 2.7405 | 3.8978 | 4.7720 |
| 0.01 | 1.1022 | 2.1871 | 4.2588 | 8.0389 | 1.2707 | 2.7453 | 6.8666 | 14.9470 |
| 0.001 | 1.1184 | 2.2242 | 4.4253 | 8.8450 | 1.1592 | 2.3886 | 4.7791 | 10.1632 |
| 0.0001 | 1.1094 | 2.2220 | 4.5413 | 8.9095 | 1.1506 | 2.2964 | 4.6193 | 9.3123 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.2389 | 2.5115 | 3.1172 | 3.7326 | 1.0023 | 1.2644 | 1.7569 | 2.4369 |
| 0.01 | 1.0702 | 2.1404 | 4.1317 | 7.7383 | 1.0042 | 1.9112 | 3.6869 | 6.7579 |
| 0.001 | 1.0957 | 2.1718 | 4.3240 | 8.5964 | 1.0081 | 1.9862 | 3.9675 | 7.8552 |
| 0.0001 | 1.0948 | 2.1779 | 4.3848 | 8.7811 | 1.0100 | 2.0006 | 3.9949 | 7.9747 |

Table 5.5: Table one of dispersion estimates $\hat{\rho}$ found using the dispersion estimation algorithm in Section 5.5.1. The fitted configurations used by the dispersion estimation algorithm are generated using the MBA approach with the power transform (4.6) using $\beta = -0.5$ and the $b_0$ value on the table; with the $m_{\text{min}} = 1$ adjustment; dispersion $\rho$ on the table and fitting into one or two dimensional space using metric MDS.

| | Straight line | | | | Parabola | | | |
|---|---|---|---|---|---|---|---|---|
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0283 | 1.7698 | 2.9020 | 4.3924 | 2.1895 | 3.3145 | 4.5449 | 6.2180 |
| 0.01 | 1.0984 | 2.1721 | 4.2158 | 8.0214 | 1.3007 | 2.8174 | 7.3471 | 15.0040 |
| 0.001 | 1.1197 | 2.2227 | 4.3894 | 8.7735 | 1.1593 | 2.3653 | 4.8158 | 9.9107 |
| 0.0001 | 1.1004 | 2.2020 | 4.4501 | 8.8886 | 1.1368 | 2.2983 | 4.6078 | 9.2423 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.1408 | 2.4918 | 3.5604 | 5.2674 | 1.0021 | 1.3231 | 2.1416 | 3.7566 |
| 0.01 | 1.0697 | 2.1009 | 4.1087 | 7.8479 | 1.0052 | 1.9157 | 3.6684 | 6.7529 |
| 0.001 | 1.0909 | 2.1703 | 4.2954 | 8.6539 | 1.0093 | 1.9827 | 3.9535 | 7.8595 |
| 0.0001 | 1.0877 | 2.2043 | 4.3779 | 8.7222 | 1.0101 | 1.9939 | 3.9890 | 8.0043 |

Table 5.6: Table two of dispersion estimates $\hat{\rho}$ found using the dispersion estimation algorithm in Section 5.5.1. The fitted configurations used by the dispersion estimation algorithm are generated using the MBA approach with the power transform (4.6) using $\beta = -0.5$ and the $b_0$ value on the table; with the $m_{\min} = 2$ adjustment; dispersion $\rho$ on the table and fitting into one or two dimensional space using metric MDS.

## 5.6 Bias correction

The unbiased simulation results, showed that $\hat{\mathbf{X}}$ from the power transform (4.6) with metric MDS, appears to be in strongest need for bias correction. There are two routes available to correct the bias, by removing the additive bias (5.19) from $\tilde{d}_{i,j}$ or shrinking $\tilde{d}_{i,j}$ to reduce the inflationary bias (5.20), to give a matrix of bias corrected perturbed distances $\tilde{\mathbf{D}}^* = (\tilde{d}^*_{i,j})$.

Correcting for the additive bias involves estimating each $\epsilon_{i,j}$ to remove it from $\tilde{d}_{i,j}$, to give $\tilde{d}^*_{i,j} = \tilde{d}_{i,j} - \epsilon_{i,j}$, where $\epsilon_{i,j}$ is the bias in $\tilde{d}_{i,j}$ described by (5.19). Removing $\epsilon_{i,j}$ could produce some $\tilde{d}^*_{i,j} < 0$ if $\epsilon_{i,j} > \tilde{d}_{i,j}$, which would have to be replaced with $\tilde{d}_{i,j}$ or $\hat{d}_{i,j}$ to retain a fittable $\tilde{\mathbf{D}}^*$. Correcting for the inflationary bias avoids the $\tilde{d}^*_{i,j} < 0$ problem.

To correct for the inflationary bias, a matrix of coefficients of inflation $\mathbf{C} = (c_{i,j})$ must be estimated. We may calculate an estimate $\hat{c}_{i,j}$ using (5.20) with $\hat{\mu}_{i,j}$ and $\hat{\rho}$ substituted for

$\mu_{i,j}$ and $\rho$, giving

$$\hat{c}_{i,j} = 1 + \frac{\beta}{2}(\beta - 1)\hat{\rho}\hat{\mu}_{i,j}^{-1}. \tag{5.26}$$

We may then obtain bias corrected perturbed distances:

$$\tilde{d}^*_{i,j} = \frac{\tilde{d}_{i,j}}{\hat{c}_{i,j}}. \tag{5.27}$$

Finally, the bias corrected perturbed distances $\tilde{\mathbf{D}}^* = (\tilde{d}^*_{i,j})$ were fitted into $p$ dimensional Euclidean space using metric MDS to obtain a corrected configuration $\hat{\mathbf{X}}^*$. The $\theta_{1:2}$ (2.12) measure of metric MDS performance; shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5), and size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) for corrected configuration, where all collected and compared with their equivalents from the MBA simulations. To recover the corrected fitted counts $\hat{\mathbf{U}}^* = (\hat{\mu}^*_{i,j})$ from $\hat{\mathbf{X}}^*$, the corrected fitted distances $\hat{\mathbf{D}}^* = (\hat{d}^*_{i,j})$ from (2.2) are first re-inflated before the transform is inverted

$$\hat{\mu}^*_{i,j} = f^{-1}(\hat{c}_{i,j}\hat{d}^*_{i,j}), \tag{5.28}$$

where $f^{-1}$ is (4.7).

### 5.6.1 Bias correction results

The bias correction simulation results for the semi-circle are plotted below and comments are made in view of the results from all the shapes, which can be found in Appendix Section D.

(a) Bias corrected $\theta_{1:2}$ values $m_{\text{min}} = 1$.

(b) Bias corrected $\theta_{1:2}$ values $m_{\text{min}} = 2$.

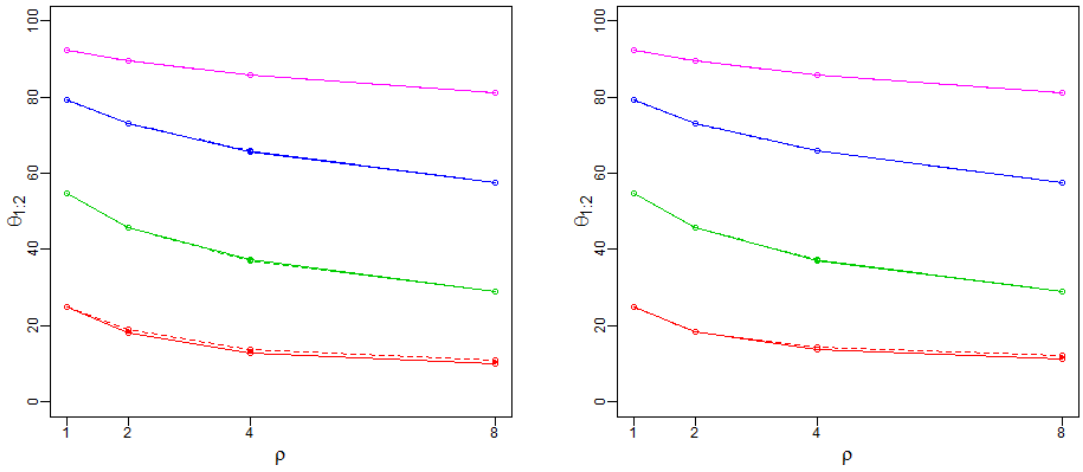



Figure 5.20: MDS performance statistics $\theta_{1:2}$ (2.12) from the bias corrected MBA simulations for a semi-circle. Left panel: $\theta_{1:2}$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ (5.27) for the power transform (4.6) with $\beta = -0.5$ and with the $m_{\text{min}} = 1$ adjustment. Right panel: $\theta_{1:2}$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ for the power transform with $\beta = -0.5$ and with the $m_{\text{min}} = 2$ adjustment. The red lines —○— gives values for $b_0 = 0.1$ (4.6); the green lines —○— for $b_0 = 0.01$; the blue lines —○— for $b_0 = 0.001$, and the pink lines —○— $b_0 = 0.0001$. The dashed lines - ⊖ - give the equivalent MBA simulation $\theta_{1:2}$ values from Figure 5.2.

## MDS performance

The bias correction has a minor decrease in $\theta_{1:p}$ compared to the MBA (uncorrected) simulations, with the decrease narrowing as the coefficient of variation $C_v(\mu_{i,j}, \rho)$ (5.4) decreases. The greatest decrease occurs on the parabola with the $m_{\text{min}} = 2$ adjustment. This could be driven by the large $\hat{\rho}$ given in Tables 5.5 and 5.6 for the parabola, over estimating the coefficient of inflation (5.26). The shrinkage is counter what is expected, ideally all the information from the perturbed distances should be projected into the first $p$ dimensions with $\theta_{1:p} = 100\%$, but perturbation disrupts information in the genuine dimensions and distributes information into the spurious dimensions. Reducing the bias shrinks $\tilde{d}_{i,j}$ but preserves the variance, so correction has reduced the bias in both genuine and spurious eigenvalues. Comparing results from setting $m_{\text{min}} = 1$ and $m_{\text{min}} = 2$, the $\theta_{1:p}$ from $m_{\text{min}} = 2$ appear to be larger, although the adjustment is only of benefit at large $C_v(\mu_{i,j}, \rho)$.

## Shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$

The bias correction produces an improvement in the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values between $\mathbf{X}$ and $\hat{\mathbf{X}}^*$ for all shapes, hence an improvement in configuration recovery. The greatest improvement is seen at large $C_v(\mu_{i,j}, \rho)$ (5.4) when the bias is most prominent. As $C_v(\mu_{i,j}, \rho)$ decreases so does the effect of the improvement. The greatest improvement is observed in the circle with the results found in Appendix Section D. The circle perturbed distance matrix is rich in large $\tilde{d}_{i,j}$ susceptible to greater perturbation. Comparing the bias correction $P(\mathbf{X}, \hat{\mathbf{X}})$ results for $m_{\text{min}} = 1$ with those for $m_{\text{min}} = 2$ in Appendix Section D, the results from $m_{\text{min}} = 2$ have the greatest improvement, although the adjustment is only of benefit at large $C_v(\mu_{i,j}, \rho)$ where the probability of simulating a $m_{i,j} = 0$ or 1 is large.

(a) Bias corrected $P(\mathbf{X}, \hat{\mathbf{X}})$ values $m_{\text{min}} = 1$.

(b) Bias corrected $P(\mathbf{X}, \hat{\mathbf{X}})$ values $m_{\text{min}} = 2$.

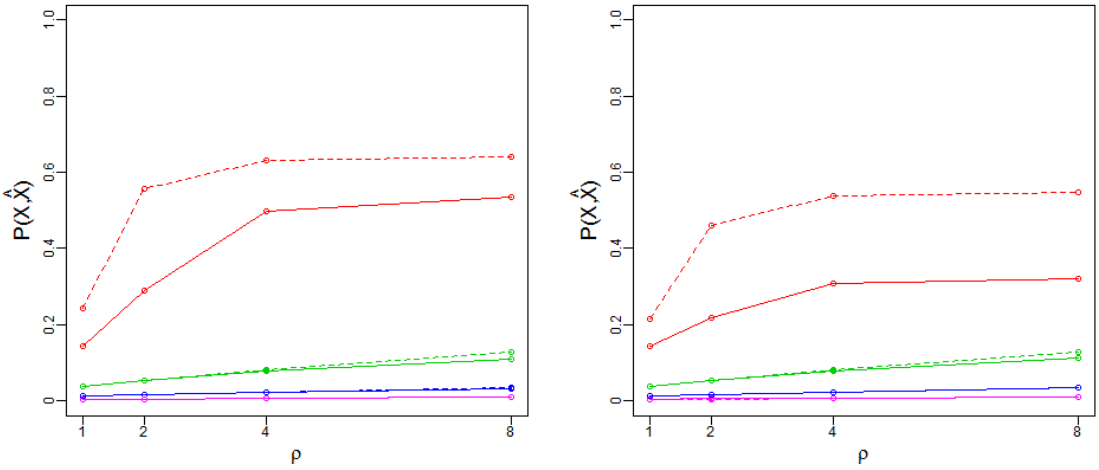



Figure 5.21: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the bias corrected MBA simulations for a semi-circle. Left panel: $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ (5.27) for the power transform (4.6) with $\beta = -0.5$ and with the $m_{\text{min}} = 1$ adjustment. Right panel: $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ or the power transform with $\beta = -0.5$ and with the $m_{\text{min}} = 2$ adjustment. The red lines —○— gives values for $b_0 = 0.1$ (4.6); the green lines —○— for $b_0 = 0.01$; the blue lines —○— for $b_0 = 0.001$, and the pink lines —○— $b_0 = 0.0001$. The dashed lines - -○- - give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3.

## Size expansion statistic $G(\mathbf{X}, \hat{\mathbf{X}})$

(a) Bias corrected $G(\mathbf{X}, \hat{\mathbf{X}})$ values $m_{\text{min}} = 1$.

(b) Bias corrected $G(\mathbf{X}, \hat{\mathbf{X}})$ values $m_{\text{min}} = 2$.

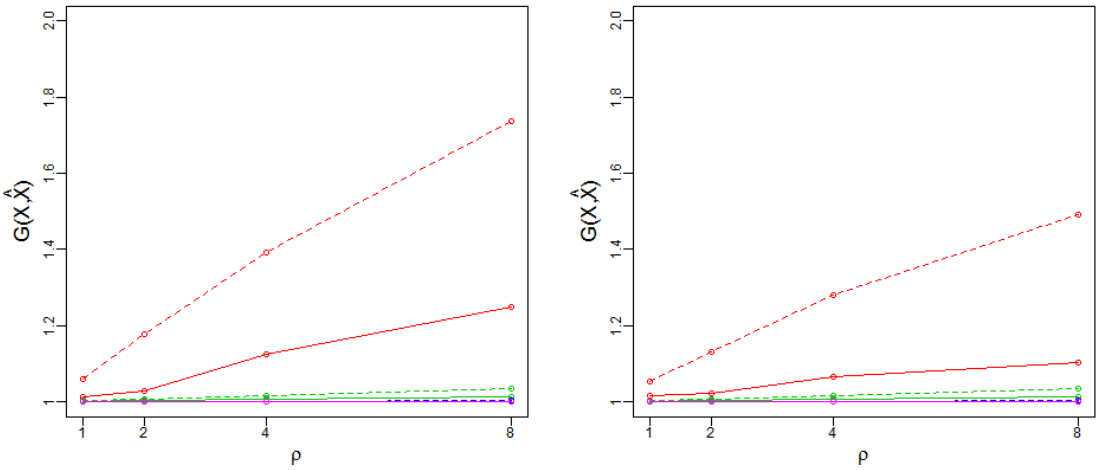



Figure 5.22: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the bias corrected MBA simulations for a semi-circle. Left panel: $G(\mathbf{X}, \hat{\mathbf{X}})$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ (5.27) for the power transform (4.6) with $\beta = -0.5$ and with the $m_{\text{min}} = 1$ adjustment. Right panel: $G(\mathbf{X}, \hat{\mathbf{X}})$ values are found using the bias corrected perturbed distances $\tilde{\mathbf{D}}^*$ for the power transform with $\beta = -0.5$ and with the $m_{\text{min}} = 2$ adjustment. The red lines —○— gives values for $b_0 = 0.1$ (4.6); the green lines —○— for $b_0 = 0.01$; the blue lines —○— for $b_0 = 0.001$, and the pink lines —○— $b_0 = 0.0001$. The dashed lines -○- give the equivalent MBA simulation $G(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.4.

The bias correction produces an improvement in $G(\mathbf{X}, \hat{\mathbf{X}})$ for all shapes, which can be observed in Figure 5.22 with the bias corrected values below their corresponding uncorrected values. The drop in $G(\mathbf{X}, \hat{\mathbf{X}})$ gives an indication of how prominent the bias is at inflating $\hat{\mathbf{X}}$. The greatest improvement is observed at large $C_v(\mu_{i,j}, \rho)$, in Figure 5.22 the margin between the corrected and uncorrected values at the $b_0$ level highlights this. The margin of improvement decreases as $C_v(\mu_{i,j}, \rho)$ (5.4) decreases. Comparing $G(\mathbf{X}, \hat{\mathbf{X}})$ results for $m_{\text{min}} = 1$ with those for $m_{\text{min}} = 2$, the results for $m_{\text{min}} = 2$ displays the greatest improvement, although the adjustment is only of benefit at large $C_v(\mu_{i,j}, \rho)$.

### Bias correction simulation summary

The bias correction results for shape difference and size expansion display the greatest improvement. The margin of improvement decreases as $C_v(\mu_{i,j}, \rho)$ decreases which is expected as less perturbation is present in the perturbed distances. Therefore bias correction provides an additional tool to improve the recovery of the original configuration $\mathbf{X}$ from the perturbed count matrix $\mathbf{M}$.

## 5.6.2 Visual comparison

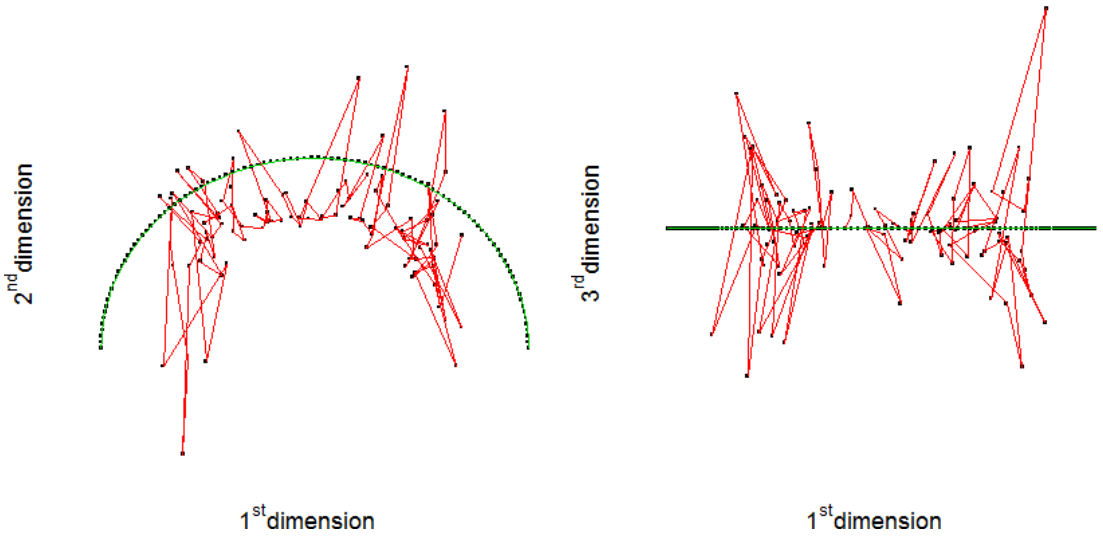



Figure 5.23: Bias corrected fitted $\hat{\mathbf{X}}^*$ and original configurations $\mathbf{X}$ for a semi-circle, generated using the MBA approach bias correction (5.27) with the power transform (4.6). The parameters for the power transform are $b_0 = 0.1$, $\beta = -0.5$; the dispersion is set at dispersion $\rho = 8$ and the $m_{\min} = 1$ adjustment is used. The matrices $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{D}}^*$ (5.27) are fitted into three dimensional Euclidean space using metric MDS. The –●– denotes a point of $\hat{\mathbf{X}}^*$ and the red line connects successive points of $\hat{\mathbf{X}}^*$. The –●– denotes a point of $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

The bias correction improves the shape difference statistic, but to gain insight into where it improves the fitted configuration visual comparison can be used. Visual comparison will plot the $\hat{\mathbf{X}}^*$ corresponding to poorest value of $P(\mathbf{X}, \hat{\mathbf{X}})$ from the power transform (4.6)

and metric MDS, generated at $b_0 = 0.1$; $\rho = 8$ and $m_{\text{min}} = 1$. In Figure 5.23 the metric MDS with the bias correction has managed to recover some structure in the second dimension, to recover $\hat{\mathbf{X}}^*$ closer to a semi-circle. Unlike the uncorrected $\hat{\mathbf{X}}$ in Figure 5.7 which has lost its semi-circular structure in the second dimension. Although noise is still present in all three dimensions and continues to inhibit the full recovery of the initial configuration from $\mathbf{M}$.

## 5.7 Application to Chromosome contact data

The bias correction can be applied to the estimated chromosome configuration for Chromosome 14 from Section 4.2.1, found using the power transform (4.6) and metric MDS $\hat{\mathbf{X}}_{P,M}$ on the chromosome count data from Hi-C (Lieberman-Aiden et al., 2009). To provide a potential improved estimate $\hat{\mathbf{X}}^*_{P,M}$.

No bias correction is applied to the estimated chromosome configurations found using the exponential transform (4.4) with metric or non-metric MDS. This is because the MBA simulation results (Appendix Section B) and unbiased simulation results (Appendix Section C) display the variance in the perturbed distances plays a larger role than the bias in the perturbed distances, in affecting shape difference. No bias correction is applied to the estimated chromosome configurations found using the power transform and non-metric MDS for similar reasons. One useful by-product of the bias correction would be an estimate of the dispersion in the Hi-C chromosome contact matrix.

To gauge if $\hat{\mathbf{X}}^*_{P,M}$ is an improvement on $\hat{\mathbf{X}}_{P,M}$, a comparison of the $\chi^2$ (4.9) was made, found using the fitted counts $\hat{\mathbf{U}}_{P.M} = (\hat{\mu}_{i,j})$ and bias corrected fitted counts $\hat{\mathbf{U}}^*_{P,M} = (\hat{\mu}^*_{i,j})$ values from (5.28). The process of applying the correction to the chromosome contact data is outlined below.

1. Apply the $m_{\text{min}}$ adjustment to $\mathbf{M}$ then the fitting algorithm from Section 4.1.3 to

obtain $\hat{\beta}$, $\hat{\mathbf{X}}_{P,M}$ (three dimensional) and $\chi^2$ (4.9). Note we set $b_0 = 1$ as $b_0$ only acts to uniformly scale distances when transforming counts to distances.

2. The fitted distances $\hat{\mathbf{D}}_{P,M}$ and fitted counts $\hat{\mathbf{U}}_{P,M}$ from $\hat{\mathbf{X}}_{P,M}$ are used to obtain $\hat{\rho}$ using the dispersion estimation algorithm in Section 5.5.1.

3. The fitted counts $\hat{\mathbf{U}}_{P,M}$ and dispersion estimate $\hat{\rho}$ are used in (5.26) to obtain the estimated coefficient of inflation $\hat{\mathbf{C}}$ and find the biased corrected distances $\tilde{\mathbf{D}}^*_{P,M}$ using (5.27).

4. The bias corrected distances $\tilde{\mathbf{D}}^*_{P,M}$ are then fitted into three dimensional Euclidean space using metric MDS to obtain the bias corrected estimated configuration $\hat{\mathbf{X}}^*_{P,M}$ and the bias corrected score function $\chi^{2*}$ (4.9).

### 5.7.1 Bias corrected estimated chromosome configuration

The bias correction was applied to Chromosome 14 to give a three dimensional bias corrected estimated chromosome configurations for Chromosome 14 $\hat{\mathbf{X}}^*_{P,M}$. With a similar analysis given to $\hat{\mathbf{X}}^*_{P,M}$, as given to $\hat{\mathbf{X}}_{P,M}$ described in Section 4.2.1. The bias correction results for each chromosomes can be found in Tables 5.9 and 5.10.

Table 5.7 shows the bias correction improves the score functions for both $m_{\text{min}}$ adjustments. The score function corresponding to the $m_{\text{min}} = 1$ adjustment produces the greatest improvement while the score functions corresponding $m_{\text{min}} = 2$ produces a smaller improvement. The shape difference between $\hat{\mathbf{X}}_{P,M}$ and $\hat{\mathbf{X}}^*_{P,M}$ was measured using POSS$(\mathbf{X}, \mathbf{Y})$ (4.10), giving POSS$(\hat{\mathbf{X}}_{P,M}, \hat{\mathbf{X}}^*_{P,M}) = 0.2389$ for the $m_{\text{min}} = 1$ adjustment and POSS$(\hat{\mathbf{X}}_{P,M}, \hat{\mathbf{X}}^*_{P,M}) = 0.0366$ for the $m_{\text{min}} = 2$ adjustment. The small change in shape for the correction corresponding to the $m_{\text{min}} = 2$ adjustment indicates the bias correction only makes a small change to the estimated chromosome configuration.

| Adjustment | $\hat{\beta}$ | $\hat{\rho}$ | $\chi^2$ | SSR$(\mathbf{M}, \hat{\mathbf{U}})$ | $S_3(\hat{\mathbf{X}})$ |
|---|---|---|---|---|---|
| $m_{\text{min}} = 1$ | -0.4497 | 2.2452 | 501915 | $3.1604\times10^9$ | 21.7088% |
| | | | (1171886) | $(2.9535\times10^{10})$ | (32.0391%) |
| $m_{\text{min}} = 2$ | -0.4796 | 2.018 | 355777 | $9.8337\times10^8$ | 19.82039 |
| | | | (485658) | $(7.4257\times10^8)$ | (23.73281%) |

Table 5.7: Score function data from using metric MDS with the bias correction for the power transform, to obtain an bias corrected (Section 5.7) estimated chromosome configuration for Chromosome 14 $\hat{\mathbf{X}}^*_{P,M}$. Column one indicates which $m_{\text{min}}$ adjustment has been used. Columns two and three give the estimated power transform (4.6) parameter value $\hat{\beta}$ (Table 4.2), and the estimated dispersion $\hat{\rho}$ found using the dispersion estimation algorithm Section 5.5.1. Column four gives the minimized $\chi^2$ (4.9) values found using the corrected fitted counts $\hat{\mathbf{U}}^*_{P,M}$ (5.28), below in brackets the uncorrected value from Table 4.2. Column five gives the SSR$(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) values found using $\hat{\mathbf{U}}^*_{P,M}$, below in brackets the uncorrected value from Table 4.2. Column six gives the $S_3(\hat{\mathbf{X}})$ (2.14) values found using $\hat{\mathbf{X}}^*_{P,M}$, below in brackets the uncorrected value from Table 4.2.

The bias correction results corresponding to the $m_{\text{min}} = 2$ adjustment will be analysed further, as this produces the lowest value of $\chi^{2*}$ (4.9).

| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_{1:3}$ |
|---|---|---|---|---|
| Corrected | 15.283% | 5.692% | 3.918% | 24.893% |
| (Uncorrected) | (16.570%) | (5.356%) | (3.875%) | (25.801%) |

Table 5.8: Percentage of information projected into the first three dimensions $\theta_1$, $\theta_2$ and $\theta_3$ and total percentage of information projected into the first three dimensions $\theta_{1:3}$. Row one: percentages of information for Chromosome 14's bias corrected estimated configuration for the power transform (4.6). Row two: percentages of information for Chromosome 14's estimated configuration for the power transform (Section 4.2.1). The bias corrected values $\theta_1$, $\theta_2, \theta_3$ and $\theta_{1:3}$ values are found by substituting the fitted eigenvalues in $\hat{\mathbf{\Lambda}}^*_{P,M}$ into (2.11) and (2.12).

In Figure 5.24 $\hat{\mathbf{X}}^*_{P,M}$ retains its less severe horseshoe shape in the first and second dimensions, and still lacks a polynomial relationship between the first and third dimensions, characteristic of horseshoe configurations. The differences between $\hat{\mathbf{X}}^*_{P,M}$ and $\hat{\mathbf{X}}_{P,M}$ appear to be more subtle, with difference seen in the local structure and points

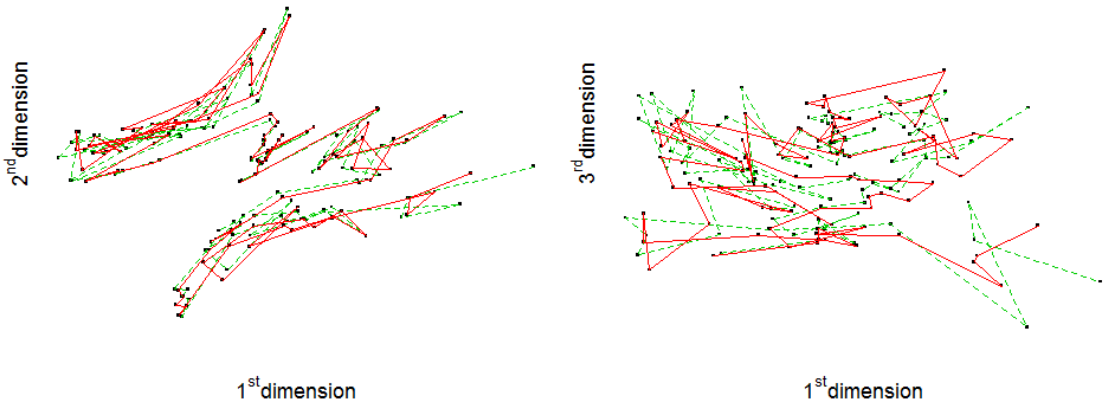



Figure 5.24: Perspectives of Chromsome 14's bias corrected estimated configuration $\hat{\mathbf{X}}^*_{P,M}$. The conifguration $\hat{\mathbf{X}}^*_{P,M}$ is found by applying the bias correction technique (Section 5.6) for the power transform to $\hat{\mathbf{X}}_{P,M}$ found in Section 4.2.1. Both distance matrices $\tilde{\mathbf{D}}$ and $\tilde{\mathbf{D}}^*$ (5.27) are fitted into three dimensional Euclidean space using metric MDS. The -●- denotes a point of $\hat{\mathbf{X}}^*_{P,M}$ and the red line connects successive points of $\hat{\mathbf{X}}^*$. The -●- denotes a point of $\hat{\mathbf{X}}_{P,M}$ and the green dashed line connects successive points of $\hat{\mathbf{X}}_{P,M}$.

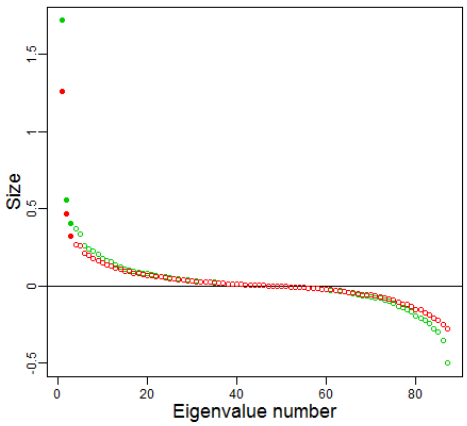



Figure 5.25: Fitted eigenvalues $\hat{\mathbf{\Lambda}}^*_{P,M}$ from fitting Chromosome 14's bias corrected estimated distances $\tilde{\mathbf{D}}^*_{P,M}$ into three dimensional Euclidean space with metric MDS. The red circles ● denote the three lead fitted eigenvalues $\hat{\lambda}^*_1, \hat{\lambda}^*_2$ and $\hat{\lambda}^*_3$ from $\hat{\mathbf{\Lambda}}^*_{P,M}$, and the hollow red circles ○ denote the remaining $\hat{\lambda}^*_k$ for $k \geq 4$. The green circles ● denote the three lead fitted eigenvalues $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_3$ from $\hat{\mathbf{\Lambda}}_{P,M}$ (Section 4.2.1), and the hollow green circles ○ denote the remaining $\hat{\lambda}_k$ for $k \geq 4$.

which jutted out from $\hat{\mathbf{X}}_{P,M}$ now slightly closer in $\hat{\mathbf{X}}^*_{P,M}$.

In Figure 5.24 the magnitude of the lead three eigenvalues has decreases, but their magnitude relative to the spurious eigenvalues appears to have improved. The magnitude of the third genuine eigenvalue is now larger than the absolute magnitude of the largest negative eigenvalues, thus satisfying the magnitude criterion and the margin from the first spurious eigenvalue appears to have increased. The spurious eigenvalues still take an "s" shape although not as pronounced as before.

Applying the bias correction to the configuration from the chromosome contact matrix, $\hat{\mathbf{X}}^*_{P,M}$ retains the overall shape of $\hat{\mathbf{X}}_{P,M}$ but subtle changes appear at a local level. The magnitude of the genuine eigenvalues improves relative to the spurious eigenvalues, although the drop in magnitude of the lead eigenvalues, causes the proportion of information projected into the lead dimensions to be less than before. In Tables 5.9 and 5.10 the bias correction produces an improved $\chi^2$ (4.9) on all the chromosomes, with the exception of Chromosome 2 using the $m_{\text{min}} = 1$ adjustment and Chromosome 6 and the X Chromosome with the $m_{\text{min}} = 2$ adjustment.

## 5.8 Conclusion

The model-based approach used simulations on known configurations to display how perturbation in the counts affects the fitted configuration. Investigating how perturbation is translated into distances, we observe that the exponential transform behaves unusually with bias and variance in the perturbed distances peaking before the boundary (at $\tilde{d}_{i,j} = 1$). The power transform behaves sensibly with the bias and variance in the perturbed distances peaking on the boundary. Unbiased simulations indicated that the power transform with metric MDS was the best candidate for a bias correction, so an estimate for dispersion was obtained and a bias correction constructed and applied. The bias correction improved the recovery of $\mathbf{X}$ in $\hat{\mathbf{X}}^*$ and performed better when $C_v(\mu_{i,j}, \rho)$

was large. The bias correction was then applied to the chromosome contact matrices and produced improved $\chi^2$ values for the majority of the chromosomes.

| Chromosome | Percentage $m_{i,j} = 0$ &1 | $\hat{\rho}$ | $\chi^2$ | $\chi^{2*}$ | $\frac{\chi^{2*}}{\chi^2}$ |
|---|---|---|---|---|---|
| 1 | 0.2045% | 1.0005 | 2179027 | 1587384 | 0.7285 |
| 2 | 0.1879% | 1.2075 | 2192977 | 2629122 | 1.1989 |
| 3 | 0.0801% | 1.1359 | 1723729 | 1313844 | 0.7622 |
| 4 | 0.1265% | 1.0958 | 1598033 | 1411594 | 0.8833 |
| 5 | 0.0909% | 1.1565 | 1450381 | 1261109 | 0.8695 |
| 6 | 0.0876% | 1.5384 | 2523582 | 1352563 | 0.5360 |
| 7 | 0.0679% | 1.4668 | 1686702 | 1520920 | 0.9017 |
| 8 | 0.0699% | 1.7270 | 1188464 | 1071292 | 0.9014 |
| 9 | 0.2342% | 2.0134 | 3747273 | 1632013 | 0.4355 |
| 10 | 0.0116% | 1.4449 | 954364 | 924317 | 0.9685 |
| 11 | 0.0239% | 2.2367 | 2258035 | 1120401 | 0.4962 |
| 12 | 0.0363% | 1.8071 | 1915093 | 660281 | 0.3448 |
| 13 | 0.0219% | 2.4995 | 948437 | 600156 | 0.6328 |
| 14 | 0.0267% | 2.4598 | 1171886 | 505551 | 0.4314 |
| 15 | 0.1543% | 1.7253 | 893767 | 787873 | 0.8815 |
| 16 | 0.0333% | 3.1847 | 1771469 | 1288205 | 0.7272 |
| 17 | 0.0333% | 2.4300 | 985024 | 670888 | 0.6811 |
| 18 | 0.0702% | 1.7468 | 424011 | 297502 | 0.7016 |
| 19 | 0.0699% | 3.3241 | 1009454 | 414690 | 0.4108 |
| 20 | 0.0584% | 3.9891 | 1051830 | 350008 | 0.3328 |
| 21 | 0.2016% | 2.9644 | 162856 | 123058 | 0.7556 |
| 22 | 0.1783% | 2.2021 | 70514 | 58464 | 0.8291 |
| X | 0.1074% | 2.9378 | 1626946 | 1228782 | 0.7553 |

Table 5.9: Summary of the bias correction technique (Section 5.7) for the power transform (4.6) with the $m_{\text{min}} = 1$ adjustment, when applied to Chromosome 1 to 22 and X's estimated configurations. Column one lists which chromosome the row of data is referring to. Column two lists percentage of elements ($m_{i,j} = 0$) in the chromosome count matrix $\mathbf{M}$ adjusted by the $m_{\text{min}} = 1$ adjustment. Column three lists the estimates for dispersion in $\mathbf{M}$ (Section 5.5.1. Column four and five list the $\chi^2$ (4.9) values before the bias correction (Table A.1,A.2 and A.3), and after the bias correction $\chi^2$ values from the Chromosomes bias corrected estimated configurations. Column six the ratio of the after and before $\chi^2$ values (the relative improvement).

| Chromosome | Percentage $m_{i,j} = 0, 1$ &2 | $\hat{\rho}$ | $\chi^2$ | $\chi^{2*}$ | $\frac{\chi^{2*}}{\chi^2}$ |
|---|---|---|---|---|---|
| 1 | 1.0423% | 1.1231 | 1445467 | 1335243 | 0.9237 |
| 2 | 0.6914% | 1.1737 | 1542520 | 1461832 | 0.9477 |
| 3 | 0.3739% | 1.1990 | 1188609 | 1026257 | 0.8634 |
| 4 | 0.7073% | 1.1774 | 1135280 | 1011563 | 0.8910 |
| 5 | 0.4026% | 1.1688 | 1023446 | 900981 | 0.8803 |
| 6 | 0.4162% | 1.2723 | 1121189 | 1175006 | 1.0480 |
| 7 | 0.2716% | 1.6573 | 1243815 | 1125652 | 0.9050 |
| 8 | 0.2198% | 1.5732 | 744585 | 678936 | 0.9118 |
| 9 | 0.8953% | 1.9175 | 1426884 | 1129849 | 0.7918 |
| 10 | 0.0347% | 1.5664 | 765634 | 804244 | 1.0504 |
| 11 | 0.2385% | 2.4250 | 1253663 | 1143319 | 0.9120 |
| 12 | 0.3270% | 1.7400 | 709648 | 684009 | 0.9639 |
| 13 | 0.1535% | 2.1229 | 738917 | 622988 | 0.8431 |
| 14 | 0.1337% | 2.0079 | 485658 | 355876 | 0.7328 |
| 15 | 0.3704% | 2.4922 | 893767 | 757224 | 0.8472 |
| 16 | 0.2331% | 3.7167 | 850103 | 614229 | 0.7225 |
| 17 | 0.0999% | 2.9400 | 677167 | 508206 | 0.7505 |
| 18 | 0.1053% | 3.1861 | 424011 | 228482 | 0.5389 |
| 19 | 0.2096% | 4.4406 | 781721 | 216780 | 0.2773 |
| 20 | 0.1169% | 3.8369 | 581958 | 339554 | 0.5835 |
| 21 | 0.6048% | 3.8788 | 162856 | 115958 | 0.7120 |
| 22 | 0.3565% | 2.3671 | 70514 | 57438 | 0.8146 |
| X | 0.4653% | 1.5711 | 650033 | 669500 | 1.0299 |

Table 5.10: Summary of the bias correction technique (Section 5.7) for the power transform (4.6) with the $m_{\text{min}} = 2$ adjustment, when applied to Chromosome 1 to 22 and X's estimated configurations. Column one lists which chromosome the row of data is referring to. Column two lists percentage of elements ($m_{i,j} = 0\&1$) in the chromosome count matrix $\mathbf{M}$ adjusted by the $m_{\text{min}} = 2$ adjustment. Column three lists the estimates for dispersion in $\mathbf{M}$ (Section 5.5.1. Column four and five list the $\chi^2$ (4.9) values before the bias correction (Table A.1,A.2 and A.3), and after the bias correction $\chi^2$ values from the Chromosomes bias corrected estimated configurations. Column six the ratio of the after and before $\chi^2$ values (the relative improvement).

# Chapter 6

# Model-based approach: extensions

Chapter 5 used a model-based approach (MBA) to detect a bias when fitting counted data into three dimensional Euclidean space, measured the biases affecting the perturbed distances and fitted configurations, and concluded with a successful bias correction technique for use with the power transform (4.6), and metric multidimensional scaling (MDS). This chapter supplements Chapter 5 by investigating biases in the fitted configurations and fitted eigenvalues, identifies the reasons why a bias correction for the exponential transform will not work, and describes novel approaches to try to correct for the bias and noise in the exponential transform.

## 6.1 Fitting the expected perturbed distances

Chapter 5 used visual comparison to interpret how perturbation was translated into the fitted configuration $\hat{\mathbf{X}} = (\hat{x}_{i,k})$. This was performed by aligning the original configuration $\mathbf{X} = (x_{i,j})$ with $\hat{\mathbf{X}}$ and plotting them together. Visual comparison gave insight into the effect of biases and noises on $\hat{\mathbf{X}}$, but noise dominated the analysis and clouded any inferences which might have been made on the effect of the bias.

To obtain a clear impression of the effect of the bias, distance matrices were constructed using the expected distances found through the delta method (5.8). This give expected distance matrices $E(\tilde{\mathbf{D}}) = (E(\tilde{d}_{i,j}))$, where each $E(\tilde{d}_{i,j})$ was constructed by inputting the original distance $d_{i,j}$ into (5.12) for the exponential transforms expected distances or (5.17) for the power transforms expected distances, the diagonal elements of $E(\tilde{\mathbf{D}})$ remained zero. The matrices $E(\tilde{\mathbf{D}})$ were fitted into three dimensional Euclidean space using metric MDS to give an expected fitted configuration $E(\hat{\mathbf{X}})$ with expected fitted eigenvalues $E(\hat{\lambda}_k)$ (2.7). Non-metric MDS was excluded from this analysis because it appeared more robust to bias in the unbiased simulations (Section 5.4). The configurations $E(\hat{\mathbf{X}})$ and $\mathbf{X}$ were aligned using the procedures used for visual comparison and plotted. To accentuate the biases the parameters which gave the largest bias were used. These parameters were $\alpha = 0.1$ for the exponential transform and $b_0 = 0.1$ and $\beta = -0.5$ for the power transform, setting the level of dispersion at $\rho = 8$ for both cases.

### 6.1.1 Exponential transform

The expected fitted configurations $E(\hat{\mathbf{X}})$ for the exponential transform (4.4) can be observed in Figures 6.1 and 6.2, the corresponding expected fitted eigenvalues $E(\hat{\lambda}_k)$ can be observed in Figures 6.5 and 6.6. In Figures 6.1 and 6.2, all shapes appear to have warped taking on characteristics of the horseshoe effect, with a horseshoe in the first and second (third for the line) dimensions and a cubic polynomial relationship in the first and third (second for the line) dimensions. The circle appears to have contracted while retaining its shape in the first and second dimensions, and gained additional structure in the third dimension to resemble a spring washer.

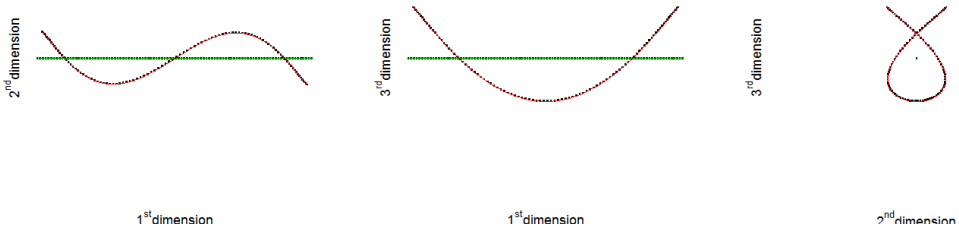
The unusual properties of the exponential transform bias (visible in Figure 5.12a) shift the middling distances into larger distance and has little effect on the larger distances, creating confusion between middling and larger distances, creating conditions contributing to the horseshoe effect.

For the circle, an abundance of $d_{i,j} = 1$ distributed uniformly between the points along with a equal distribution of all the distances, buttress the circle against the bias. The equal distribution of distances means that each point in $\mathbf{X}$ shares the same set of distances between the other points.

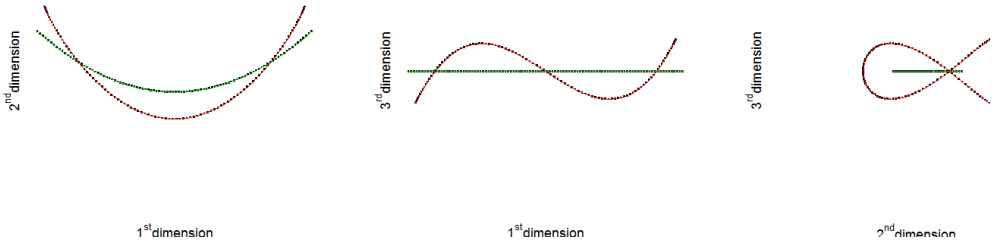
In Figures 6.5 and 6.6, for the straight line, parabola and semi-circle, the magnitude of the genuine $E(\hat{\lambda}_k)$ is larger than the genuine $\lambda_k$ for $k = 1, 2$. This increase is driven by the bias, increasing the length of the distances. For the circle, $E(\hat{\lambda}_k) < \lambda_k$ for $k = 1, 2$, due to the buttressing effect limiting the size of $E(\hat{\lambda}_k)$ and forcing information into the spurious dimensions. The bias produces spurious $E(\hat{\lambda}_k)$ for $k \geq 2$ with a small number of $E(\hat{\lambda}_k) > 0$ and the remaining $E(\hat{\lambda}_k) = 0$. No spurious $E(\hat{\lambda}_k) < 0$ are produced, suggesting $E(\tilde{\mathbf{D}})$ from the exponential transform does not violate the Euclidean properties of a distance matrix. The absence of negative expected eigenvalues $E(\hat{\lambda}_k)$ suggests the fitted eigenvalues $\hat{\lambda}_k$ in Figures 4.4 and 5.6 are the sole product of the noise in $\tilde{\mathbf{D}}$.

### 6.1.2 Power transform

The expected fitted configurations $E(\hat{\mathbf{X}})$ for the power transform (4.6) can be observed in Figures 6.3 and 6.4, the corresponding expected fitted eigenvalues $E(\hat{\lambda}_k)$ can be observed in Figure 6.7 and 6.8. In Figures 6.3 and 6.4, the bias appears to have stretched the shapes and in some cases stretched information out of the first two dimensions. The line, parabola and semi-circle are related shapes, having successively more information in the second dimension. The bias stretches the line and parabola into a one dimensional shape. The semi-circle retains information in the second dimension and gains information in the third dimension. The circle remains robust to the bias and expands in a similar manner to when the expected perturbed distances $E(\tilde{\mathbf{D}})$ from when the exponential transform are fitted into Euclidean space, to produce a expected fitted configuration resembling a spring washer.




(a) $E(\hat{\mathbf{X}})$ for the straight line.

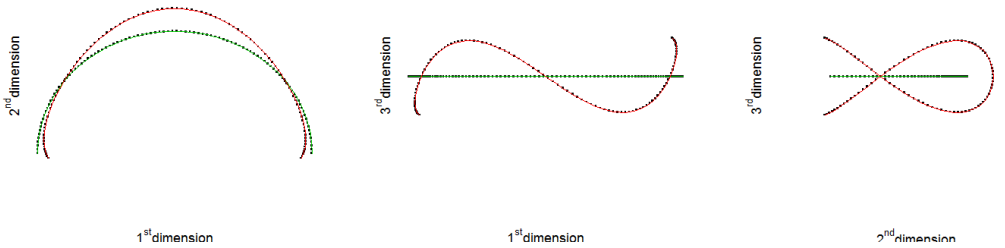



(b) $E(\hat{\mathbf{X}})$ for the parabola.

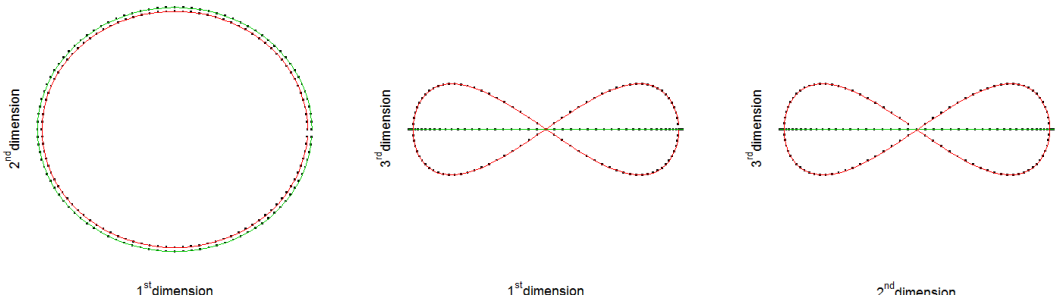
Figure 6.1: Expected configurations $E(\hat{\mathbf{X}})$, from fitting the MBA expected exponential transform distances $E(\tilde{\mathbf{D}})$ (5.12) with metric MDS. Row one: $E(\hat{\mathbf{X}})$ for a straight line. Row two: $E(\hat{\mathbf{X}})$ for a parabola. Each $E(\tilde{\mathbf{D}})$ is generated using $\alpha = 0.1$ and $\rho = 8$. The ●– denotes a point of $E(\hat{\mathbf{X}})$ and the red line connects successive points of $E(\hat{\mathbf{X}})$. The ●– denotes a point of the original configuration $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

The horseshoe effect appears less prominent in $E(\hat{\mathbf{X}})$ from the power transform. The size of the bias increases as the distance increases, so small distances remain small distances; medium distances remain medium distances and large distances remain large distances, lacking the confusion between medium and large distances which contributes to the horseshoe effect.

In Figures 6.7 and 6.8 for the straight line; parabola and semi-circle, $E(\hat{\lambda}_1) > \lambda_1$ and for the parabola and semi-circle $E(\hat{\lambda}_2) < \lambda_2$, indicating information is lost in the second dimension. For the circle, $E(\hat{\lambda}_k) > \lambda_k$ for $k = 1, 2$. For the semi-circle and circle, bias has produced some spurious $E(\hat{\lambda}_k) > 0$ for $k > 2$. In all shapes, the bias produces




(a) $E(\hat{\mathbf{X}})$ for the semi-circle.



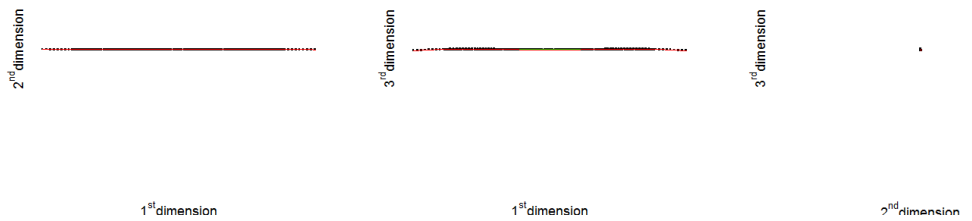

(b) $E(\hat{\mathbf{X}})$ for the circle.

Figure 6.2: Expected configurations $E(\hat{\mathbf{X}})$, from fitting the MBA expected exponential transform distances $E(\tilde{\mathbf{D}})$ (5.12) with metric MDS. Row one: $E(\hat{\mathbf{X}})$ for a semi-circle. Row two: $E(\hat{\mathbf{X}})$ for a circle. Each $E(\tilde{\mathbf{D}})$ is generated using $\alpha = 0.1$ and $\rho = 8$. The ●– denotes a point of $E(\hat{\mathbf{X}})$ and the red line connects successive points of $E(\hat{\mathbf{X}})$. The ●– denotes a point of the original configuration $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

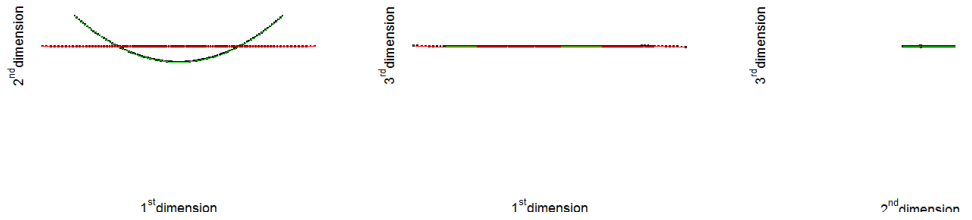
spurious $E(\hat{\lambda}_k) < 0$ for $k \geq 96$, to compensate for the majority of information being projected into the first dimension. Hence $E(\tilde{\mathbf{D}})$ violates the Euclidean properties of a distance matrix. The negative fitted eigenvalues $\hat{\lambda}_k$ in Figures 4.4 and 5.8 are a product of bias and noise in $\tilde{\mathbf{D}}$; this is why the bias correction reduces the magnitude of the negative spurious eigenvalues.

### 6.1.3 Conclusion

The exponential and power transform have different biassing characteristics in $E(\hat{\mathbf{X}})$ and $E(\hat{\lambda})$. The exponential bias (5.14) warps the shapes and utilizes additional dimensions



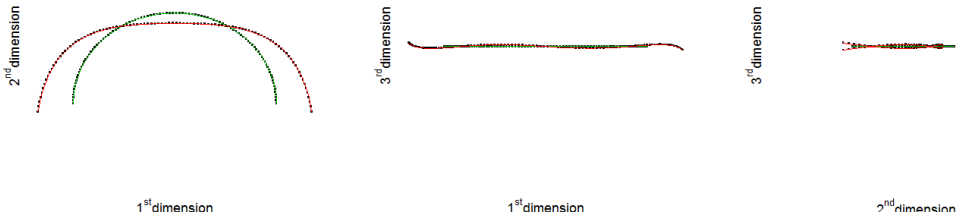

(a) $E(\hat{\mathbf{X}})$ for the straight line.



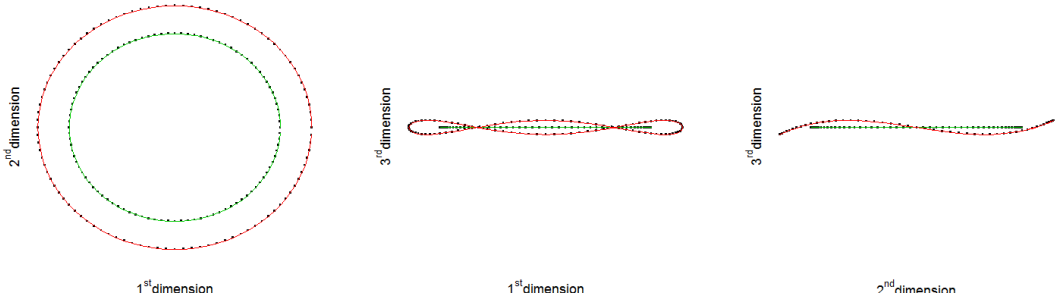

(b) $E(\hat{\mathbf{X}})$ for the parabola.

Figure 6.3: Expected configurations $E(\hat{\mathbf{X}})$, from fitting the MBA expected power transform distances $E(\tilde{\mathbf{D}})$ (5.17) with metric MDS. Row one: $E(\hat{\mathbf{X}})$ for a straight line. Row two: $E(\hat{\mathbf{X}})$ for a parabola. Each $E(\tilde{\mathbf{D}})$ is generated using $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$. The ●– denotes a point of $E(\hat{\mathbf{X}})$ and the red line connects successive points of $E(\hat{\mathbf{X}})$. The ●– denotes a point of the original configuration $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

to accommodate the bias, middling distances are increased to large distances while large distances are increased only a little, causing the confusion that forces $E(\hat{\mathbf{X}})$ to take on aspects of the horseshoe effect, although the exponential transform $E(\tilde{\mathbf{D}})$ retains its Euclidean properties. The power transform bias (5.19) stretches shapes which are not robust to the bias, drawing information out of the second dimension and into the first. The power transform bias increases as distance increases avoiding the confusion between middling and large distances, making the horseshoe effect less prominent in $E(\hat{\mathbf{X}})$, although the power transforms $E(\tilde{\mathbf{D}})$ loses its Euclidean properties and requires additional complex dimensions to absorb the increase in information in the lead dimension. Studies similar to using $E(\tilde{\mathbf{D}})$ and $E(\hat{\mathbf{X}})$ were undertaken using Kato



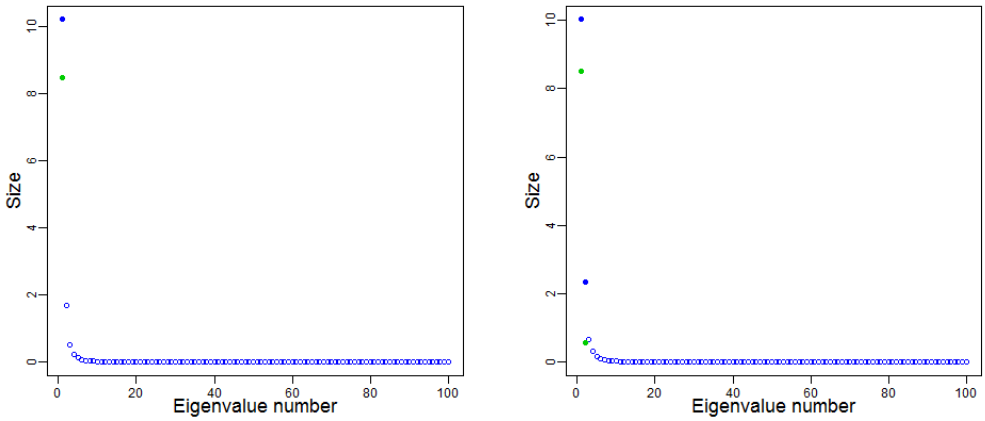

$E(\hat{\mathbf{X}})$ for the semi-circle.




$E(\hat{\mathbf{X}})$ for the circle.

Figure 6.4: Expected configurations $E(\hat{\mathbf{X}})$, from fitting the MBA expected power transform distances $E(\tilde{\mathbf{D}})$ (5.17) with metric MDS. Row one: $E(\hat{\mathbf{X}})$ for a semi-circle. Row two: $E(\hat{\mathbf{X}})$ for a circle. Each $E(\tilde{\mathbf{D}})$ is generated using $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$. The ●– denotes a point of $E(\hat{\mathbf{X}})$ and the red line connects successive points of $E(\hat{\mathbf{X}})$. The ●– denotes a point of the original configuration $\mathbf{X}$ and the green line connects successive points of $\mathbf{X}$.

approximation (Kato, 1966; Sibson, 1979; Kent et al., 1983). The Kato approximation provided a linear approximation to the biases effect on the eigenvalues and eigenvectors. This was used to plot how the bias moved the points in the fitted configuration. The Kato approximation found for the exponential transform the bias promoted aspects of the horseshoe effect in the fitted configuration, while for the power transform the bias had more of a stretching effect.



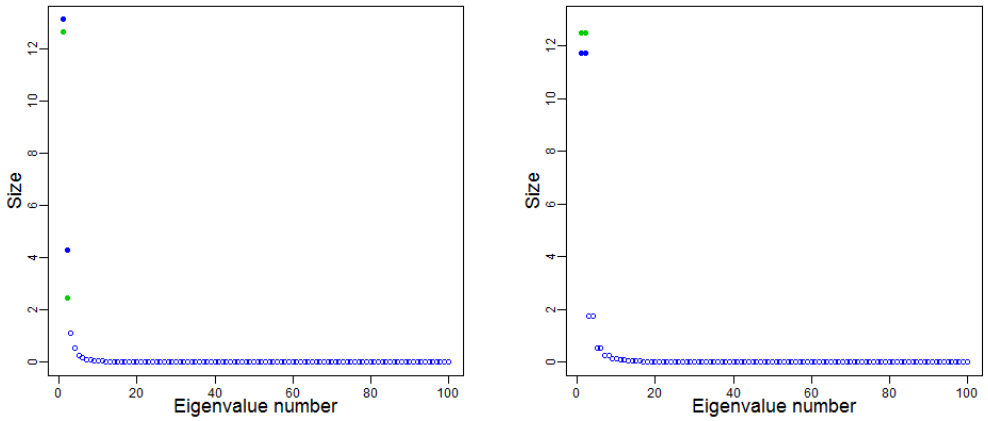

(a) $E(\hat{\lambda}_k)$ for the straight line. (b) $E(\hat{\lambda}_k)$ for the parabola.

Figure 6.5: Expected fitted eigenvalues $E(\hat{\lambda}_k)$ (2.7), from fitting the MBA expected exponential transform distances $E(\tilde{\mathbf{D}})$ (5.12) with metric MDS. Left panel: the $E(\hat{\lambda}_k)$ for a straight line. Right panel: the $E(\hat{\lambda}_k)$ for a parabola. Each $E(\tilde{\mathbf{D}})$ is generated using $\alpha = 0.1$ and $\rho = 8$. The blue circles ● denote genuine $E(\hat{\lambda}_k)$ for $k \leq 2$; the hollow blue circles ○ denote spurious $E(\hat{\lambda}_k)$ for $k \geq 2$. The green circles ● denote the original eigenvalues $\lambda_k$ for $k \leq 2$.

## 6.2 Trials

Chapter 5 identified a bias in both count to distance transform functions and provided a correction technique when using power transform (4.6) with metric MDS. Four trials were run to detect if a similar bias correction was possible for the exponential transform (4.4). The trial displaying the greatest improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) might then be used as a bias correction technique with real Hi-C count data.

The bias correction for the power transform described in Section 5.6 used the fitted counts $\hat{\mathbf{U}} = (\hat{\mu}_{i,j})$ and an estimate of dispersion $\hat{\rho}$ to produce a matrix of estimated inflation coefficients $\hat{\mathbf{C}} = (\hat{c}_{i,j})$. The estimated inflation coefficients $\hat{\mathbf{C}}$ then corrected $\tilde{\mathbf{D}}$ to produce




(a) $E(\hat{\lambda}_k)$ for the semi-circle. (b) $E(\hat{\lambda}_k)$ for the circle.

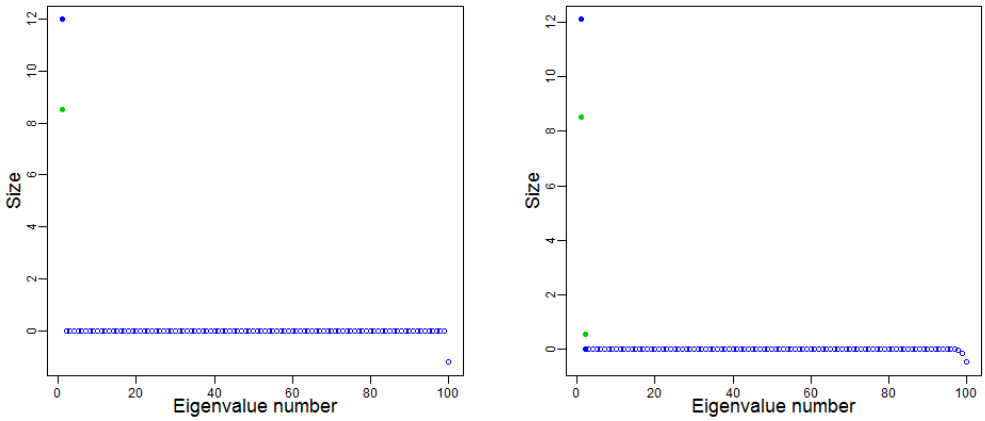
Figure 6.6: Expected fitted eigenvalues $E(\hat{\lambda}_k)$ (2.7), from fitting the MBA expected exponential transform distances $E(\tilde{\mathbf{D}})$ (5.12) with metric MDS. Left panel: the $E(\hat{\lambda}_k)$ for a semi-circle. Right panel: the $E(\hat{\lambda}_k)$ for a circle. Each $E(\tilde{\mathbf{D}})$ is generated using $\alpha = 0.1$ and $\rho = 8$. The blue circles ● denote genuine $E(\hat{\lambda}_k)$ for $k \leq 2$; the hollow blue circles ○ denote spurious $E(\hat{\lambda}_k)$ for $k \geq 3$. The green circles ● denote the original eigenvalues $\lambda_k$ for $k \leq 2$.

a bias corrected perturbed distance matrix $\tilde{\mathbf{D}}^* = (\tilde{d}^*_{i,j})$ (5.27) where

$$\tilde{d}^*_{i,j} = \frac{\tilde{d}_{i,j}}{\hat{c}_{i,j}}.$$

Finally $\tilde{\mathbf{D}}^*$ was fit into Euclidean space using metric MDS to obtain a bias-corrected fitted configuration $\hat{\mathbf{X}}^*$.

The trials reported in this section for the exponential transform work on a similar basis to the bias correction for the power transform. Each trial used a different expression for the coefficient of inflation $\mathbf{C} = (c_{i,j})$ calculated from the original distances $d_{i,j}$, allowing the search for a coefficient of inflation which offers the most promise as a correction. Using the true values of $c_{i,j}$ in the trials, and not estimates, removes any inaccuracy which could undermine the success of the trial. If no trials were successful using $c_{i,j}$ then a correction using $\hat{c}_{i,j}$ would be even less likely.

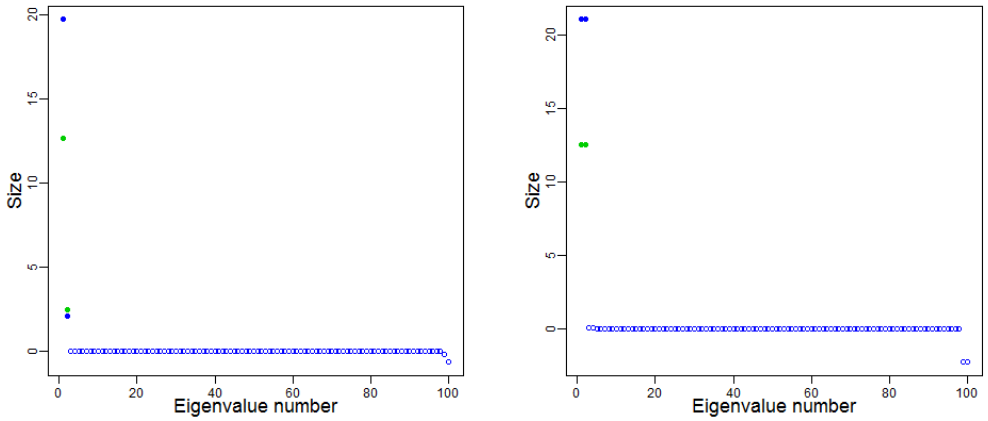



(a) $E(\hat{\lambda}_k)$ for the straight line.

(b) $E(\hat{\lambda}_k)$ for the parabola.

Figure 6.7: Expected fitted eigenvalues $E(\hat{\lambda}_k)$ (2.7), from fitting the MBA expected power transform distances $E(\tilde{\mathbf{D}})$ (5.17) with metric MDS. Left panel: the $E(\hat{\lambda}_k)$ for a straight line. Right panel: the $E(\hat{\lambda}_k)$ for a parabola. Each $E(\tilde{\mathbf{D}})$ is generated using $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$. The blue circles ● denote genuine $E(\hat{\lambda}_k)$ for $k \leq 2$; the hollow blue circles ○ denote spurious $E(\hat{\lambda}_k)$ for $k \geq 2$. The green circles ● denote the original eigenvalues $\lambda_k$ for $k \leq 2$.

Stage One of the trials used $\mathbf{C}$ to bias correct the $\tilde{\mathbf{D}}$ using (5.27) to produce $\tilde{\mathbf{D}}^*$, which was then fitted into Euclidean space using metric MDS to give $\hat{\mathbf{X}}^*$. The corresponding shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between $\mathbf{X}$ and $\hat{\mathbf{X}}^*$ from the trial was then calculated and compared with the original $P(\mathbf{X}, \hat{\mathbf{X}})$ from the MBA simulation. Stage One was repeated 1000 times to provide a mean value for $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). The trial providing the largest improvement in shape difference was then used in Stage Two. Stage Two used $\hat{c}_{i,j}$ calculated using $\hat{d}_{i,j}$ and $\rho$. If the shape difference improved in Stage Two then Stage Three would use $\hat{c}_{i,j}$ calculated using $\hat{d}_{i,j}$ and $\hat{\rho}$ similar to the power-transform bias correction (5.26). The trials were tested on $\tilde{\mathbf{D}}$ from a parabola with dispersion $\rho = 8$, these conditions providing the largest $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) in the MBA simulations, therefore providing scope for a correction to improve $P(\mathbf{X}, \hat{\mathbf{X}})$. To differentiate between the trials, additional subscript notation is added to the matrices, for example $\mathbf{C}_1 = (c_{1,i,j})$ is the coefficient of inflation for Trial 1.




(a) $E(\hat{\lambda}_k)$ for the semi-circle (b) $E(\hat{\lambda}_k)$ for the circle.

Figure 6.8: Expected fitted eigenvalues $E(\hat{\lambda}_k)$ (2.7), from fitting the MBA expected power transform distances $E(\tilde{\mathbf{D}})$ (5.17) with metric MDS. Left panel: the $E(\hat{\lambda}_k)$ for a semi-circle. Right panel: the $E(\hat{\lambda}_k)$ for a circle. Each $E(\tilde{\mathbf{D}})$ is generated using $b_0 = 0.1$, $\beta = -0.5$ and $\rho = 8$. The blue circles ● denote genuine $E(\hat{\lambda}_k)$ for $k \leq 2$; the hollow blue circles ○ denote spurious $E(\hat{\lambda}_k)$ for $k \geq 3$. The green circles ● denote the original eigenvalues $\lambda_k$ for $k \leq 2$.

### 6.2.1 Constructing the trials

Four different expressions for $c_{i,j}$ were found, using the delta method approximation to the expectations of the perturbed distances $f(m_{i,j}) = \tilde{d}_{i,j}$, and the delta method approximation to the expectation of the perturbed squared distances $g(m_{i,j}) = \tilde{d}^2_{i,j}$.

Trials 1 and 2 used the delta-method approximation to the expectation of $f(m_{i,j})$. The Taylor-series expansion of $f(m_{i,j})$ around $\mu_{i,j}$ to third order was taken

$$\begin{aligned} f(m_{i,j}) \approx\ & e^{-\alpha\mu_{i,j}} - \alpha e^{-\alpha\mu_{i,j}}(m_{i,j} - \mu_{i,j}) \\ & + \frac{\alpha^2}{2} e^{-\alpha\mu_{i,j}}(m_{i,j} - \mu_{i,j})^2 - \frac{\alpha^3}{6} e^{-\alpha\mu_{i,j}}(m_{i,j} - \mu_{i,j})^3 \end{aligned} \tag{6.1}$$

Then taking the expectation of (6.1), gives the delta method approximation to the expectations of $f(m_{i,j})$ to third order

$$E(f(m_{i,j})) \approx d_{i,j} - \frac{\alpha}{2}\log(d_{i,j})d_{i,j}\rho + \frac{\alpha^2}{6}\log(d_{i,j})d_{i,j}\rho(2\rho-1). \tag{6.2}$$

Trial 1 used the coefficient of inflation extracted from the first two terms of (6.2)

$$c_{1,i,j} \approx 1 - \frac{\alpha}{2}\log(d_{i,j})\rho,$$

which is the same as (5.15). Trial 2 used the coefficient of inflation extracted from the full expression of (6.2)

$$c_{2,i,j} = 1 - \frac{\alpha}{2}\log(d_{i,j})\rho + \frac{\alpha^2}{6}\log(d_{i,j})\rho(2\rho-1).$$

Trials 3 and 4 used the delta-method approximation to the expectation of $g(m_{i,j}) = e^{-2\alpha m_{i,j}}$. The Taylor-series expansion of $g(m_{i,j})$ around $\mu_{i,j}$ to third order was taken

$$\begin{aligned} g(m_{i,j}) \approx\ & e^{-2\alpha\mu_{i,j}} - 2\alpha e^{-2\alpha\mu_{i,j}}(m_{i,j}-\mu_{i,j}) + \frac{4}{2!}\alpha^2 e^{-2\alpha\mu_{i,j}}(m_{i,j}-\mu_{i,j})^2 \\ & - \frac{8}{3!}\alpha^3 e^{-2\alpha\mu_{i,j}}(m_{i,j}-\mu_{i,j})^3. \end{aligned} \tag{6.3}$$

Then taking the expectation of (6.3), gives the delta method approximation to the expectations of $g(m_{i,j})$ to third order

$$E(g(m_{i,j})) \approx d_{i,j}^2 - 2\alpha d_{i,j}^2\log(d_{i,j})\rho + \frac{4}{3}\alpha^2 d_{i,j}^2\log(d_{i,j})\rho(2\rho-1). \tag{6.4}$$

Trial 3 used the coefficient of inflation extracted from the first two terms of (6.4)

$$c_{3,i,j} = 1 - 2\alpha\log(d_{i,j})\rho.$$

Trial 4 used the coefficient of inflation extracted from the full expression of (6.4)

$$c_{4_{i,j}} = 1 - 2\alpha\log(d_{i,j})\rho + \frac{4}{3}\alpha^2\log(d_{i,j})\rho(2\rho - 1).$$

Trials 1 and 2 corrected $\tilde{\mathbf{D}}$ in a way similar to the power transform bias correction in Section 5.6, and Trials 3 and 4 corrected the perturbed squared distances $\tilde{\mathbf{D}}^2$ to give $\tilde{\mathbf{D}}_t^{2*} = \tilde{d}_{t,i,j}^{2*}$

$$\tilde{d}_{t,i,j}^{2*} = \frac{\tilde{d}_{i,j}^2}{c_{t,i,j}} \text{ where } t = 3 \text{ or } 4.$$

### 6.2.2 Trial results

The shape difference statistic $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) for the trials is displayed in Table 6.1

| | | Stage one | | | | Stage two |
|---|---|---|---|---|---|---|
| $\alpha$ | Original | Trial 1 | Trial 2 | Trial 3 | Trial 4 | Trial 2 |
| 0.1 | 0.3495 | 0.3734 | 0.3091 | 0.4286 | 0.3494 | 0.3485 |
| 0.01 | 0.1962 | 0.2148 | 0.2133 | 0.2538 | 0.2450 | 0.2171 |
| 0.001 | 0.0708 | 0.0712 | 0.0712 | 0.0724 | 0.0720 | 0.0717 |
| 0.0001 | 0.0225 | 0.0224 | 0.0224 | 0.0225 | 0.0224 | 0.0226 |

Table 6.1: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the MBA bias correction trials for the exponential transform (4.4). The trials were run to bias correction for the perturbed distances for a parabola with dispersion $\rho = 8$. Column one lists the levels of $\alpha$ the rows correspond to. Column two the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA simulations (Table B.9). Column three to six the $P(\mathbf{X}, \hat{\mathbf{X}})$ values for Trials 1,2,3 and 4 stage one. Column seven the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Trials 2 stage two.

The only significant improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) in Table 6.1 at Stage one occurs in Trial 2 for $\alpha = 0.1$, displaying around a $10\%$ improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$. Other insignificant improvements in Table 6.1 at Stage One occurred in Trials 1, 2 and 4 at $\alpha = 0.0001$, which amounts to a $0.5\%$ improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$. The remaining results in Table 6.1 at Stage One show a poorer $P(\mathbf{X}, \hat{\mathbf{X}})$. The Trials produce a poorer $\hat{\mathbf{X}}^*$ most

of the time, and are detrimental to the recovery of $\mathbf{X}$ from $\mathbf{M}$. Trial 2 only managed to show a strong improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). Therefore Trial 2 was brought forward for testing at Stage Two (using the fitted distances $\hat{d}_{i,j}$ and true dispersion $\rho$).

Table 6.1 shows that Stage Two Trial 2 only marginally improved $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) at $\alpha = 0.1$ and on all other levels of $\alpha$ was detrimental to $P(\mathbf{X}, \hat{\mathbf{X}})$. Therefore Trial 2 was not brought forward for testing at stage three.

### 6.2.3 Conclusion

These trials demonstrate that a successful pre-processing bias correction for the exponential transform is not easy to produce, with any pre-processing technique detrimental to the recovery of the original shape. However, the principle of using trials to find the coefficient of inflation giving the greatest bias correction could be applied to the power transform to improve the current bias correction.

## 6.3 Fit-and-Correct approach

The bias correction used information garnered through the delta method to pre-process the perturbed distances $\tilde{\mathbf{D}}$, proving successful for the power transform (4.6) and unsuccessful for the exponential transform (4.4). An alternative approach to finding a bias correction for the exponential transform involves a second round of perturbation and fitting into Euclidean space to observe how the fitted counts and fitted distances alter, then by feeding the information observed from the second round back into the original perturbed counts or distances in the form of a correction. We call this the Fit-and-Correct approach. Figure 6.9 provides a schematic of the Fit-and-Correct approach. Fit-and-Correct provides many different bias-correction techniques, as there are many ways to interpret the changes after the second round of perturbation and fitting and many ways to feed that information back

into the original perturbed data. Here one example of the Fit-and-Correct approach will be given.

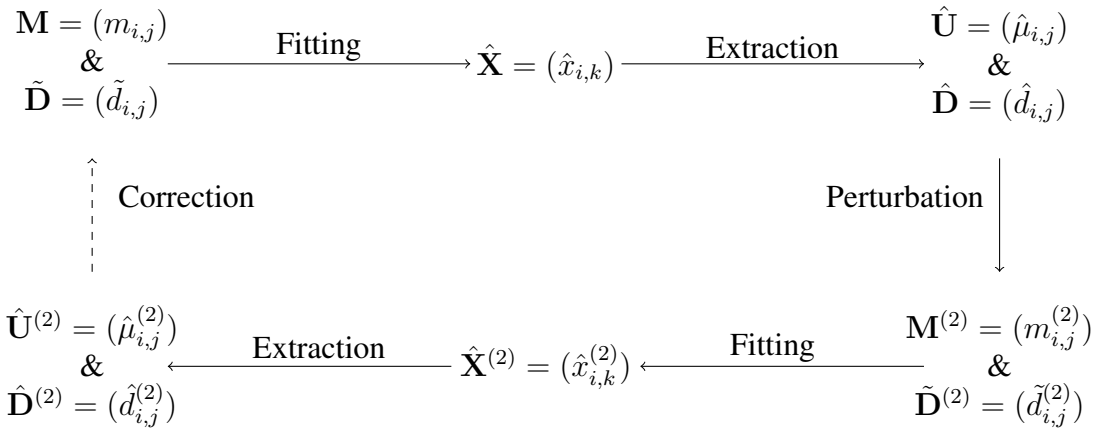



Figure 6.9: Schematic of the Fit-and-Correct technique.

### 6.3.1 Fit-and-Correct example

This example uses perturbed counts $\mathbf{M}$ from a parabola with dispersion $\rho = 8$, transformed into perturbed distances $\tilde{\mathbf{D}}$ using the exponential transform (4.5). The perturbed distance matrix $\tilde{\mathbf{D}}$ was first fitted into two-dimensional Euclidean space using metric MDS to give $\hat{\mathbf{X}}$. Then $\hat{\mathbf{D}}$ and $\hat{\mathbf{U}}$ were extracted from $\hat{\mathbf{X}}$ using (2.2) and (4.5). A second round of perturbation was applied using $\hat{\mathbf{U}}$ as the matrix of mean counts

$$m_{i,j}^{(2)} \sim \text{NB}(\frac{\hat{\mu}_{i,j}}{8-1}, \frac{1}{8})$$

and $m_{i,j}^{(2)} = m_{j,i}^{(2)}$ for symmetry. The new perturbed count matrix $\mathbf{M}^{(2)}$ was then transformed into $\tilde{\mathbf{D}}^{(2)}$ using the same transform function and parameters used to transform $\mathbf{M}$ into $\tilde{\mathbf{D}}$, and fitted into two-dimensional Euclidean space using metric MDS to give $\hat{\mathbf{X}}^{(2)}$. Then $\hat{\mathbf{D}}^{(2)}$ and $\hat{\mathbf{U}}^{(2)}$ were extracted from $\hat{\mathbf{X}}^{(2)}$ using (2.2) and (4.5). Using the new perturbed distance matrix $\tilde{\mathbf{D}}^{(2)}$ and new fitted distance matrix $\hat{\mathbf{D}}^{(2)}$, a matrix of adjustment

coefficients $\mathbf{\Phi} = (\phi_{i,j})$ was found, where

$$\phi_{i,j} = \frac{\hat{d}_{i,j}^{(2)}}{\tilde{d}_{i,j}^{(2)}},$$

which measures the extent to which the MDS fitting, inflated the distances. This process was repeated $100$ times to build an three dimensional array, where each plate of the array was an individual $\mathbf{\Phi}$, the mean was taken over the plates of the array to give mean valued matrix of adjustment coefficients $\overline{\mathbf{\Phi}}$. This gave an empirical description in $\overline{\mathbf{\Phi}}$ of how fitting adjusts the perturbed distances. The mean valued matrix of adjustment coefficients $\overline{\mathbf{\Phi}}$ is unitless, which means when applied to another matrix it does not alter the units of the matrix. The $\overline{\mathbf{\Phi}}$ was used to pre-process $\tilde{\mathbf{D}}$ to counter any adjustment made in the first round of fitting, to give $\tilde{\mathbf{D}}^* = (\tilde{d}_{i,j}^*)$ where

$$\tilde{d}_{i,j}^* = \frac{\tilde{d}_{i,j}}{\phi_{i,j}}.$$

The distance matrix $\tilde{\mathbf{D}}^*$ was then fitted into two-dimensional Euclidean space using metric MDS to obtain a corrected configuration $\hat{\mathbf{X}}^*$, and the measures of fit used in the MBA were collected and compared with the original measures of fit. In Table 6.2 the

| $\alpha$ | Original | Adjusted |
|---|---|---|
| $0.1$ | 0.3495 | 0.4198 |
| $0.01$ | 0.1962 | 0.2669 |
| $0.001$ | 0.0708 | 0.0741 |
| $0.0001$ | 0.0225 | 0.0226 |

Table 6.2: Shape difference values from the Fit-and-Correct example. The Fit-and-Correct was run to provide an improved fitted configuration from the perturbed distances for a parabola with dispersion $\rho = 8$. Column one lists the levels of $\alpha$ the rows correspond to. Column two the $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the MBA simulations (Table B.9). Column three gives the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from using the Fit-and-Correct approach.

correction deteriorates the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) and therefore is detrimental to

the recovery of the original configuration.

### 6.3.2 Conclusion

Several different approaches at trying to remove bias or adjust the perturbed distances $\tilde{\mathbf{D}}$ to improve the recovery of the original configuration were attempted, all proving unsuccessful. Fit-and-Correct contains several weaknesses preventing it becoming a successful correction technique. The first weakness is it a recursion of the MBA. The MBA assumes $\mathbf{X}$ is unknown so a correction technique can be found to recover a $\hat{\mathbf{X}}^*$ without having to rely on unknown values, but the $\hat{\mathbf{X}}^*$ can be compared with the original $\mathbf{X}$ to gauge the corrections effect. Now the Fit-and-Correct is making the $\hat{\mathbf{X}}$ become the unknown, which is unnecessary. The second weakness is that $\hat{\mathbf{U}}$ contains noise from the original perturbation and the fitting, a second round of perturbation only adds another layer of noise making $\mathbf{M}^{(2)}$ even more dissimilar to $\mathbf{U}$. The third weakness is that the second round of perturbation is also adding a second layer of bias to the new perturbed distances.

## 6.4 Post-processing with smoothing splines

Post-processing involves the uses of splines to smooth out the noise in $\hat{\mathbf{X}}$ from metric multidimensional scaling (MDS). This involves finding a balance on how much smoothing should be applied, between smoothing out the noise from $\hat{\mathbf{X}}$ and avoiding erasing features of the original configuration $\mathbf{X}$ recovered in $\hat{\mathbf{X}}$. Here we describe a method that combines smoothing splines and the delta-method estimate for the variance in perturbed distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$ to find this balance.

### 6.4.1 Splines

Splines are piecewise-polynomial functions (Green and Silverman (1994)) which fit to the data points $(\hat{x}_{i,k}, t_i)$, where $\hat{x}_{i,k}$ and $t_i$ are the dependent and explanatory variables respectively, $i = 1, \ldots, n$ and $k = 1, 2$ is the dimension number. The spline used to smooth noise out of $\hat{\mathbf{X}}$ is a smoothing spline which tries to fit a piece-wise cubic function $P_k(t_i) = y_{i,k}$ to the data according to some roughness penalty parameter $\tau > 0$. The function $P(t_i)$ tries to minimize the penalized sum of squares

$$T(P_k) = \sum_{i=1}^{n} (\hat{x}_{i,k} - y_{i,k})^2 + \tau \int_{t_0}^{t_n} (P_k''(t))^2 dt,$$

where the first term is the goodness-of-fit term and the second term is a roughness penalty. If $\tau \to 0$ the spline interpolates the points (under smoothed) and if $\tau \to \infty$ the spline tends to the linear least squares solution (over smoothed). The parameter $\tau$ will be used to control the level of smoothing the spline applies to $\hat{\mathbf{X}}$. The spline will be used to smooth each dimension of $\hat{\mathbf{X}}$ independently with the same value of $\tau$ therefore the same level of smoothing, to give a smoothed fitted configuration $\hat{\hat{\mathbf{X}}} = (\hat{\hat{x}}_{i,k})$.

### 6.4.2 The smoothing algorithm

To determine the level of smoothing to apply to $\hat{\mathbf{X}}$ a distance based indicator $w(\tilde{\mathbf{D}})$ was developed which used the percentage error between two different estimates for the total variance in $\tilde{\mathbf{D}}$. The first variance estimate was an empirical estimate for the total variance in $\tilde{\mathbf{D}}$:

$$\tilde{v}(\tilde{\mathbf{D}}) = \sum_{i<j} (\tilde{d}_{i,j} - \hat{d}_{i,j})^2. \tag{6.5}$$

The fitted distances $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ are used in place of the original distances $\mathbf{D} = (d_{i,j})$, as the $\mathbf{D}$ are unavailable. The second variance estimate used the sum of the delta-method approximations for the variance in the perturbed distances var$(\tilde{d}_{i,j})$ (5.9):

$$\hat{v}(\tilde{\mathbf{D}}) = \sum_{i<j} \text{var}(\tilde{d}_{i,j}), \tag{6.6}$$

found using $\hat{\mathbf{D}} = (\hat{d}_{i,j})$ instead of $\mathbf{D} = (d_{i,j})$. Both $\tilde{v}(\tilde{\mathbf{D}})$ (6.5) and $\hat{v}(\tilde{\mathbf{D}})$ (6.6) use $\hat{\mathbf{D}}$ instead of $\mathbf{D}$, this substitution constricts the post-processing to using $\hat{\mathbf{X}}$ from metric MDS as scale in the distances is preserved. Then the percentage error between $\tilde{v}(\tilde{\mathbf{D}})$ and $\hat{v}(\tilde{\mathbf{D}})$ is found to give the variance score:

$$w(\tilde{\mathbf{D}}) = \frac{|\tilde{v}(\tilde{\mathbf{D}}) - \hat{v}(\tilde{\mathbf{D}})|}{\tilde{v}(\tilde{\mathbf{D}})} \times 100\%. \tag{6.7}$$

The idea behind $w(\tilde{\mathbf{D}})$ is that if $\tilde{v}(\tilde{\mathbf{D}})$ and $\hat{v}(\tilde{\mathbf{D}})$ were calculated using the original distances then $w(\tilde{\mathbf{D}})$ should be small, whereas using the fitted distances to calculate $w(\tilde{\mathbf{D}})$ introduces extra error making the value large. The splines smooth noise out of $\hat{\mathbf{X}}$ this should move the values of $\hat{\mathbf{D}}$ closer to $\mathbf{D}$, correcting some of the error and reducing the size of $w(\tilde{\mathbf{D}})$ (6.7). The smoothing algorithm is described below:

1. Using the fitted distance $\hat{\mathbf{D}}$ and the perturbed distances $\tilde{\mathbf{D}}$ (from the MBA) calculate $w(\tilde{\mathbf{D}})$ (6.7) for the fitted configuration $\hat{\mathbf{X}}$, found using metric MDS.

2. Starting at some initial smoothing parameter value of $\tau = 0$, independently smooth over the $p$ dimensions of $\hat{\mathbf{X}}$ to obtain a smoothed fitted configuration $\hat{\hat{\mathbf{X}}} = (\hat{\hat{x}}_{i,k})$. Extract the fitted distances from $\hat{\hat{\mathbf{X}}}$ using (2.2) and recalculate $w(\tilde{\mathbf{D}})$ (6.7).

3. Scan across the interval $\tau = (0, \infty]$, smoothing $\hat{\mathbf{X}}$ to obtain $\hat{\hat{\mathbf{X}}}$ and collecting the corresponding $w(\tilde{\mathbf{D}})$ (6.7) values.

4. Choose the $\hat{\hat{\mathbf{X}}}$ which gives the minimized $w(\tilde{\mathbf{D}})$ value.

### 6.4.3 Post-processing for the exponential transform

The delta method estimate of the variance (5.13) in the perturbed distances from the exponential transform (4.4) is

$$\text{var}(\tilde{d}_{i,j}) \approx -\alpha\rho\log(d_{i,j})d_{i,j}^2,$$

where $\alpha$ is the parameter used by the exponential transform to determine the relationship between distances and counts, and $\rho$ is the level of dispersion used in perturbing the counts, the $\alpha$ and $\rho$ parameters are used in the MBA to determine the levels of perturbation translated into $\hat{\mathbf{X}}$. Substituting the fitted distances in (5.13) for the original distances and summing over variance estimates for the upper triangle of the distance matrix, gives the second variance estimate $\hat{v}(\tilde{\mathbf{D}})$ (6.6) for the exponential transform

$$\hat{v}(\tilde{\mathbf{D}}) = -\alpha\rho\sum_{i<j}\log(\hat{d}_{i,j})\hat{d}_{i,j}^2. \tag{6.8}$$

The smoothing algorithm was applied to $\hat{\mathbf{X}}$ from the MBA generated using the exponential transform and metric MDS. This was repeated as a simulation, to obtain a more robust insight into the smoothing algorithms performance. How the smoothing algorithm was run as a simulation for the exponential transform is described below.

1. Using the same process used in the MBA (Chapter 5) for the exponential transform and metric MDS, choose an original configuration $\mathbf{X}$, a level of $\alpha$ and dispersion $\rho$, then generate $\hat{\mathbf{X}}$.

2. Apply the smoothing algorithm to $\hat{\mathbf{X}}$ to obtain $\hat{\hat{\mathbf{X}}}$. Collect the shape difference statistics $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between $\mathbf{X}$ and $\hat{\hat{\mathbf{X}}}$ and the minimized $w(\tilde{\mathbf{D}})$ (6.7) value.

3. Repeat steps 1. and 2. for 1000 times with the same $\mathbf{X}$, $\alpha$ and $\rho$, collecting the $P(\mathbf{X}, \hat{\mathbf{X}})$ and $w(\tilde{\mathbf{D}})$ values, then find the mean of the $P(\mathbf{X}, \hat{\mathbf{X}})$ and $w(\tilde{\mathbf{D}})$ values.

The simulations were run on each of the original configurations (line; parabola; semi-circle and circle), on each of the levels of $\alpha$ used in the MBA simulations ($\alpha = 0.1$; 0.01; 0.001 and 0.0001) and each of the four levels of dispersion used in the MBA simulations ($\rho = 1$; 2; 4 and 8).

**Simulation results**

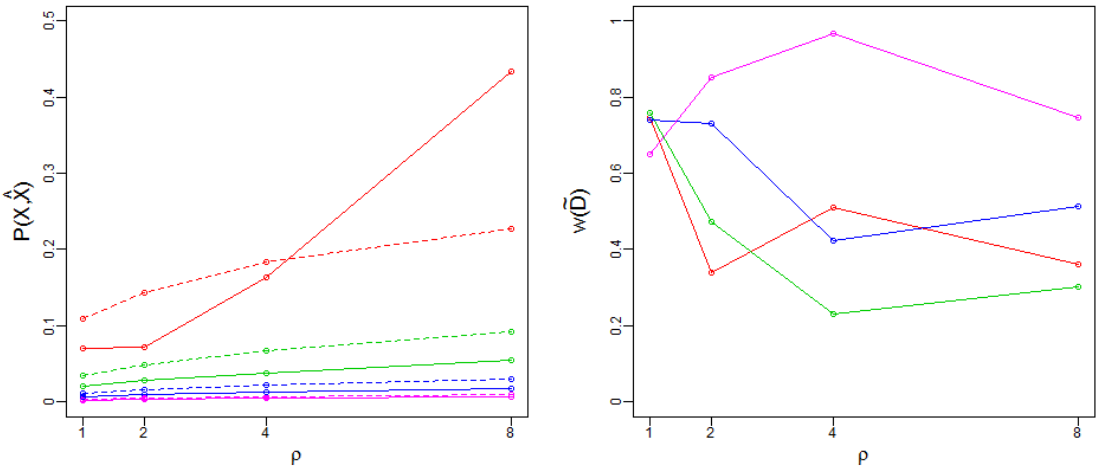
The simulation results when smoothing algorithm was applied to the semi-circle are displayed in Figure 6.10 and the simulation results for the remaining shapes can be found in Appendix Section E.

**Shape difference statistic** $P(\mathbf{X}, \hat{\mathbf{X}})$

In $\hat{\mathbf{X}}$ from the straight line and parabola at each level of $\alpha$ and $\rho$, the smoothing algorithm improved $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). The largest improvement is observed in $\hat{\mathbf{X}}$ for the straight line generated with $\alpha = 0.1$ and $\rho = 1$, here the smoothing algorithm reduces the $P(\mathbf{X}, \hat{\mathbf{X}})$ value by approximately 77.22%. Although the smoothing algorithm fails by causing a deterioration in the $P(\mathbf{X}, \hat{\mathbf{X}})$ value, for the semi-circle at $\alpha = 0.1$ and $\rho = 8$ and the circle at $\alpha = 0.1$ and $\rho \geq 2$ or $\alpha = 0.01$ and $\rho = 8$.

**Variance score** $w(\tilde{\mathbf{D}})$

The variance score statistic $w(\tilde{\mathbf{D}})$ (6.7) is not as easy to interpret as $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). The $w(\tilde{\mathbf{D}})$ values corresponding to where the smoothing algorithm fails are not unusually larger, than the $w(\tilde{\mathbf{D}})$ values for when the smoothing algorithm is successful. The size of the $w(\tilde{\mathbf{D}})$ values vary between the shapes. There is no relationship between the $w(\tilde{\mathbf{D}})$ values and the coefficient of variation $\mathrm{C}_v(\mu_{i,j}, \rho)$ (5.4). Figure 6.10b displays how varied they can be, with values from $\alpha = 0.0001$ producing some of the largest $w(\tilde{\mathbf{D}})$ values.

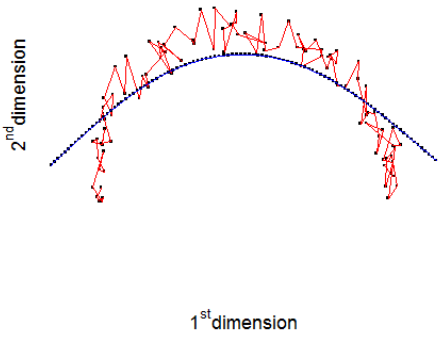



(a) $P(\mathbf{X}, \hat{\mathbf{X}})$. (b) $w(\tilde{\mathbf{D}})$.

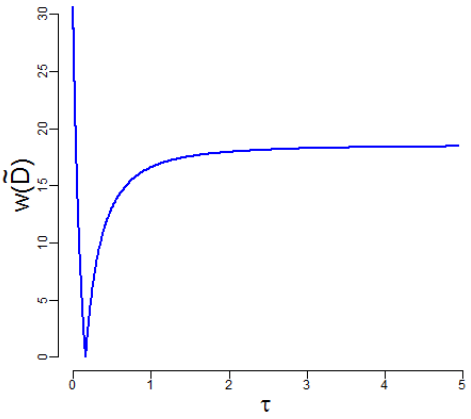
Figure 6.10: Simulation results from using smoothing splines to post-process the fitted configuration $\hat{\mathbf{X}}$, to give the fitted smoothed configuration $\hat{\hat{\mathbf{X}}}$. The configuration $\hat{\mathbf{X}}$ is generated from a semi-circle; using the exponential transform (4.4) and metric MDS. The configuration $\hat{\hat{\mathbf{X}}}$ is generated by applying the smoothing algorithm (Section 6.4.2) for the exponential transform to $\hat{\mathbf{X}}$. Left panel: shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between $\mathbf{X}$ and $\hat{\hat{\mathbf{X}}}$. Right panel: variance score $w(\tilde{\mathbf{D}})$ (6.7) values. The red lines —○— gives values for $\alpha = 0.1$ (4.4); the green lines —○— for $\alpha = 0.01$; the blue lines —○— for $\alpha = 0.001$, and the pink lines —○— $\alpha = 0.0001$. The dashed lines - -○- - in the left panel give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3.

**Visual comparison**

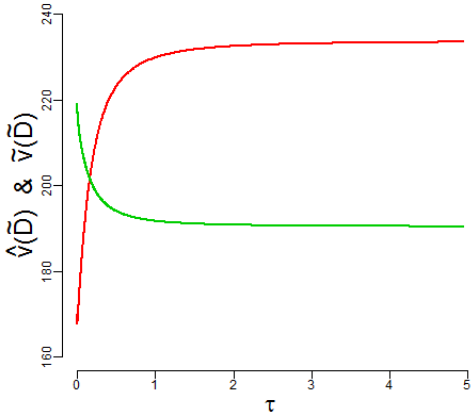
Figure 6.11a displays the $\hat{\mathbf{X}}$ and $\hat{\hat{\mathbf{X}}}$ for a semi-circle generated using $\alpha = 0.1$ and dispersion $\rho = 4$, (the parameters giving the largest level of $\mathrm{C}_v(\mu_{i,j}, \rho)$ (5.4) before the smoothing algorithm fails). The smoothing algorithm produces $\hat{\hat{\mathbf{X}}}$ which resembles more of a parabola than a semi-circle. This parabola shape is an indication of over-smoothing and although the simulation produced and improved $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) the improvement was only marginal. The $\tilde{v}(\tilde{\mathbf{D}})$ (6.5) and $\hat{v}(\tilde{\mathbf{D}})$ (6.8) values neatly meet at $\tau = 0.1358$ giving a $w(\tilde{\mathbf{D}}) = 0$. As $\tau$ increases beyond $\tau > 0.1358$ the $\tilde{v}(\tilde{\mathbf{D}})$, $\hat{v}(\tilde{\mathbf{D}})$ and $w(\tilde{\mathbf{D}})$ (6.7) values become constant which suggests the shape of $\hat{\hat{\mathbf{X}}}$ changes little with increased smoothing. The failure of the smoothing in the semi-circle and circle, could be occurring




(a) $\hat{\mathbf{X}}$ and $\hat{\hat{\mathbf{X}}}$ from the semi-circle.




(b) $w(\tilde{\mathbf{D}})$.




(c) $\tilde{v}(\tilde{\mathbf{D}})$ and $\hat{v}(\tilde{\mathbf{D}})$.

Figure 6.11: Individual result from post-processing the fitted configuration $\hat{\mathbf{X}}$, to give a fitted smoothed configuration $\hat{\hat{\mathbf{X}}}$. The configuration $\hat{\mathbf{X}}$ is generated from a semi-circle using the exponential transform (4.4) with $\alpha = 0.1$; with dispersion set at $\rho = 4$ and using metric MDS. The configuration $\hat{\hat{\mathbf{X}}}$ is generated by applying the smoothing algorithm (Section 6.4.2) for the exponential transform to $\hat{\mathbf{X}}$. The variance score $w(\tilde{\mathbf{D}})$ (6.7) has a minimized value of $w(\tilde{\mathbf{D}}) = 0$ at $\tau = 0.1358$. Top: configurations $\hat{\mathbf{X}}$ and $\hat{\hat{\mathbf{X}}}$. The —● denotes a point of $\hat{\mathbf{X}}$ and the red line connects successive points of $\hat{\mathbf{X}}$. The —● denotes a point of $\hat{\hat{\mathbf{X}}}$ and the blue line connects successive points of $\hat{\hat{\mathbf{X}}}$. Bottom left: the variance score $w(\tilde{\mathbf{D}})$ denoted by the blue line — as the smoothing parameter $\tau$ increases. Bottom right: variance estimates $\tilde{v}(\tilde{\mathbf{D}})$ (6.5) denoted by the red line —, and $\hat{v}(\tilde{\mathbf{D}})$ (6.8) denoted by the green line — as $\tau$ increases.

for the parabola and straight line and are passing undetected, as these shapes are used by the spline $P_k(t_i)$. For example the spline could over-smooth the parabola, by fitting a least squares cubic function to the data, with emphasis on the quadratic term so $\hat{\tilde{\mathbf{X}}}$ still resembles a parabola.

### 6.4.4 Post-processing for the power transform

The post-processing for the power transform (4.6) can be applied to the bias-corrected configuration $\hat{\mathbf{X}}^*$, to give a pre-processed (for bias) and post-processed (for noise) configuration $\hat{\tilde{\mathbf{X}}}^*$. The bias correction corrects the perturbed distances using (5.27) to give bias corrected perturbed distances $\tilde{\mathbf{D}}^* = (\tilde{d}^*_{i,j})$. To acknowledge the bias correction the estimates $\tilde{v}(\tilde{\mathbf{D}})$ (6.5) and $\hat{v}(\tilde{\mathbf{D}})$ (6.6) require adjustment to estimate the total variance in $\tilde{\mathbf{D}}^*$. In $\tilde{v}(\tilde{\mathbf{D}}^*)$ the $\tilde{\mathbf{D}}^*$ are used in place of $\tilde{\mathbf{D}}$ to give

$$\tilde{v}(\tilde{\mathbf{D}}^*) = \sum_{i<j} (\tilde{d}^*_{i,j} - \hat{d}_{i,j})^2. \tag{6.9}$$

The delta-method estimate for the variance in the bias corrected perturbed distances is

$$\text{var}(\tilde{d}^*_{i,j}) = b_0 \beta^2 \rho \frac{d_{i,j}^{2-\frac{1}{\beta}}}{\hat{c}^2_{i,j}},$$

this gives a delta method estimate for total variance of

$$\hat{v}(\tilde{\mathbf{D}}) = b_0 \beta^2 \rho \sum_{i<j} \frac{\hat{d}_{i,j}^{2-\frac{1}{\beta}}}{\hat{c}^2_{i,j}} \tag{6.10}$$

The smoothing algorithm was applied to the bias corrected fitted configuration $\hat{\mathbf{X}}^*$ from the MBA, generated using the power transform (4.6) with $m_{\text{min}} = 2$ adjustment, fitted in to Euclidean space with metric MDS and corrected using the bias correction (5.27).

This was repeated in a simulation, to obtain a more robust insight into the smoothing algorithms performance. How the smoothing algorithm was run as a simulation for the power transform is described below.

1. Using the same process in the MBA for the power transform and metric MDS, using an original configuration $\mathbf{X}$, a level of $b_0$ with $\beta = -0.5$, a level of dispersion $\rho$ and the $m_{\text{min}} = 2$ adjustment, generate $\hat{\mathbf{X}}$.

2. Apply the bias correction technique for the power transform to obtain $\hat{\mathbf{C}}$, $\tilde{\mathbf{D}}^*$ and $\hat{\mathbf{X}}^*$. Using the true $\rho$ not the estimated $\rho$.

3. Apply the smoothing algorithm to $\hat{\mathbf{X}}^*$ to obtain the bias corrected smoothed fitted configuration $\hat{\hat{\mathbf{X}}}^*$. Collect the shape difference statistics $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between $\mathbf{X}$ and $\hat{\hat{\mathbf{X}}}^*$ and the minimized $w(\tilde{\mathbf{D}})$ (6.7) value.

4. Repeat steps 1. 2. and 3. for 1000 times with the same $\mathbf{X}$; $b_0$; $\beta$; $\rho$ and $m_{\text{min}} = 2$, collecting the $P(\mathbf{X}, \hat{\mathbf{X}})$ values and $w(\tilde{\mathbf{D}})$ values, then find the mean of the $P(\mathbf{X}, \hat{\mathbf{X}})$ and $w(\tilde{\mathbf{D}})$ values.

The simulations were run each of the four original configurations (line; parabola; semi-circle and circle), on each of the levels of $b_0$ used in the MBA simulations ($b_0 = 0.1$; 0.01; 0.001 and 0.0001) with $\beta = -0.5$ and on each of the levels of dispersion used in the MBA simulations ($\rho = 1$; 2; 4 and 8), with $m_{\text{min}} = 2$.

**Simulation results**

The simulation results when smoothing algorithm was applied to the semi-circle are displayed in Figure 6.12 and the simulation results for the remaining shapes can be found in Appendix Section E.

**Shape difference statistic** $P(\mathbf{X}, \hat{\mathbf{X}})$

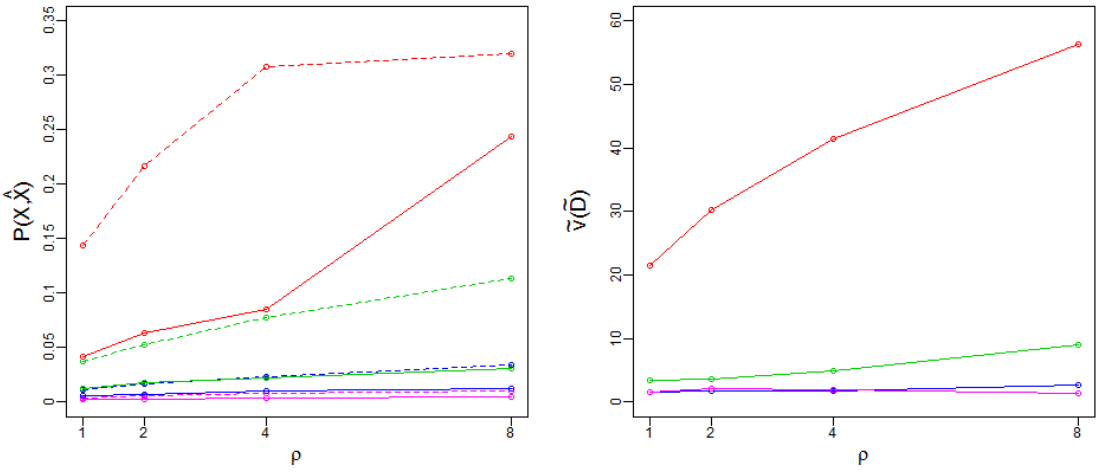
In each the simulations the smoothing algorithm has managed to improve $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). The largest improvement is observed when the smoothing algorithm is applied to $\hat{\mathbf{X}}^*$ from a straight line, generated using the power transform with $b_0 = 0.1$, $\beta = -0.5$ and dispersion $\rho = 4$, here the smoothing algorithm reduces $P(\mathbf{X}, \hat{\mathbf{X}})$ by approximately 95.37%. The mean reduction in $P(\mathbf{X}, \hat{\mathbf{X}})$ from the application of the smoothing algorithm to $\hat{\mathbf{X}}^*$ is approximately 61.09%. These results from the smoothing algorithm simulations indicate it is a useful tool in the recovery of the original configuration from the perturbed count matrix.

**Variance score** $w(\tilde{\mathbf{D}})$

The variance score $w(\tilde{\mathbf{D}})$ (6.7) for the power transform simulations appears easier to interpret that the equivalent for the exponential transform simulations. The value of $w(\tilde{\mathbf{D}})$ appears to increase as the coefficient of variation $\mathrm{C}_v(\mu_{i,j}, \rho)$ (5.4) increases, evidence that the smoothing algorithm is working for the power transform. The $\tau$ values at the minimized $w(\tilde{\mathbf{D}})$ are usually very small ($\tau \approx 0$), suggesting a little smoothing and more interpolation has been applied.

**Visual comparison**

Figure 6.13a displays $\hat{\mathbf{X}}^*$ and $\hat{\hat{\mathbf{X}}}^*$ for a semi-circle generated using $b_0 = 0.1$, $\beta = -0.5$ with dispersion $\rho = 8$ and $m_{\text{min}} = 2$. The points of $\hat{\hat{\mathbf{X}}}^*$ appear to follow the average path of the points of $\hat{\mathbf{X}}^*$, giving a smoother configuration. The global structure of $\hat{\hat{\mathbf{X}}}^*$ still resembles a semi-circle. The local structure of $\hat{\hat{\mathbf{X}}}^*$ is much less chaotic with adjacent points much closer together. The variance estimators $\tilde{v}(\tilde{\mathbf{D}}^*)$ (6.9) and $\hat{\tilde{v}}(\tilde{\mathbf{D}}^*)$ (6.10) in Figure 6.13c never meet on the range of $\tau$ scanned across and the minimized $w(\tilde{\mathbf{D}}^*)$ is



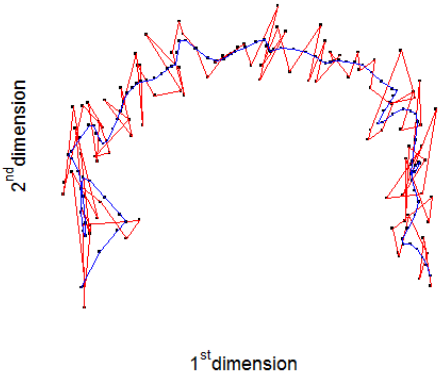

(a) $P(\mathbf{X}, \hat{\mathbf{X}})$. (b) $w(\tilde{\mathbf{D}})$.

Figure 6.12: Simulation results from using smoothing splines to post-process the bias corrected fitted configuration $\hat{\mathbf{X}}^*$, to give the bias corrected smoothed fitted configuration $\hat{\hat{\mathbf{X}}}^*$. The configuration $\hat{\mathbf{X}}^*$ is generated from a semi-circle; using the power transform (4.4) with $\beta = -0.5$; with the $m_{\min} = 2$ adjustment; metric MDS and the bias correction (Section 5.6). The configuration $\hat{\hat{\mathbf{X}}}^*$ is generated by applying the smoothing algorithm (Section 6.4.2) for the power transform to $\hat{\mathbf{X}}^*$. Left panel: shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between $\mathbf{X}$ and $\hat{\hat{\mathbf{X}}}^*$. Right panel: variance score $w(\tilde{\mathbf{D}})$ (6.7) values. The red lines —○— gives values for $b_0 = 0.1$ (4.6); the green lines —○— for $b_0 = 0.01$; the blue lines —○— for $b_0 = 0.001$, and the pink lines —○— $b_0 = 0.0001$. The dashed lines -○- in the left panel give the equivalent bias corrected MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.21b.

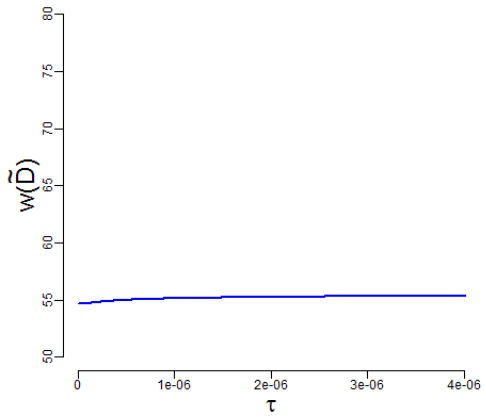
found at $\tau = 3.1152 \times 10^{-8}$. This $\tau$ value and $\hat{\hat{\mathbf{X}}}^*$ shows the smoothing algorithm was not overzealous in smoothing noise out from $\hat{\hat{\mathbf{X}}}^*$.

### 6.4.5 Application to the estimated chromosome configuration

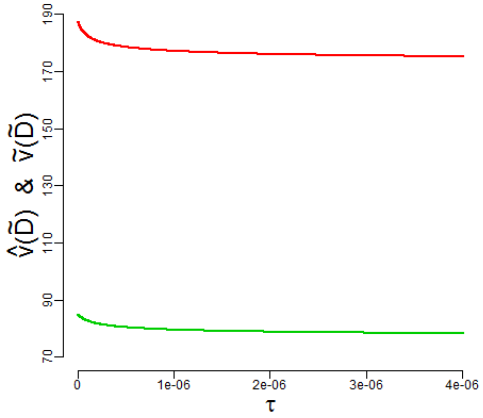
The smoothing algorithm proved successful when applied to $\hat{\mathbf{X}}^*$ from the power transform (4.6), which means it could be applied to $\hat{\mathbf{X}}^*_{P,M}$ (from Section 5.7) for chromosome 14 from the Hi-C (Lieberman-Aiden et al., 2009) count data. This should




(a) $\hat{\mathbf{X}}^*$ and $\hat{\hat{\mathbf{X}}}$ from the semi-circle.




(b) $w(\tilde{\mathbf{D}})$.




(c) $\tilde{v}(\tilde{\mathbf{D}}^*)$ and $\hat{v}(\tilde{\mathbf{D}}^*)$.

Figure 6.13: Individual result from post-processing the bias corrected fitted configuration $\hat{\mathbf{X}}^*$, to give a bias corrected fitted smoothed configuration $\hat{\hat{\mathbf{X}}}^*$. The configuration $\hat{\mathbf{X}}^*$ is generated from a semi-circle using the power transform (4.6) with $b_0 = 0.1$ and $\beta = -0.5$; with dispersion set at $\rho = 8$; with the $m_{\min} = 2$ adjustment; using metric MDS and the bias correction (Section 5.6). The configuration $\hat{\hat{\mathbf{X}}}^*$ is generated by applying the smoothing algorithm (Section 6.4.2) for the power transform to $\hat{\mathbf{X}}^*$. The variance score $w(\tilde{\mathbf{D}})$ (6.7) has a minimized value of $w(\tilde{\mathbf{D}}) = 54.6499$ at $\tau = 3.1152 \times 10^{-8}$. Top: configurations $\hat{\mathbf{X}}^*$ and $\hat{\hat{\mathbf{X}}}^*$. The —● denotes a point of $\hat{\mathbf{X}}^*$ and the red line connects successive points of $\hat{\mathbf{X}}^*$. The —● denotes a point of $\hat{\hat{\mathbf{X}}}^*$ and the blue line connects successive points of $\hat{\hat{\mathbf{X}}}^*$. Bottom left: $w(\tilde{\mathbf{D}})$ values denoted by the blue line — as the smoothing parameter $\tau$ increases. Bottom right: variance estimates $\tilde{v}(\tilde{\mathbf{D}})$ (6.9) denoted by the red line —, and $\hat{v}(\tilde{\mathbf{D}})$ (6.10) denoted by the green line — as $\tau$ increases.

provide a pre-processed (for bias) and post-processed (for noise) $\hat{\hat{\mathbf{X}}}^*_{P,M}$ if successful.

The smoothing algorithm was applied to the chromosome 14's bias corrected estimated chromosome configuration $\hat{\hat{\mathbf{X}}}^*_{P,M}$ using the level of dispersion $\hat{\rho} = 2.018$ used for the bias correction. Figure 6.14 gives the bias corrected and post-processed $\hat{\hat{\mathbf{X}}}^*_{P,M}$ for chromosome 14 from the Hi-C count data. The smoothing algorithm appears to have made minor changes at a local scale with $\hat{\hat{\mathbf{X}}}^*_{P,M}$ still resembling $\hat{\mathbf{X}}^*_{P,M}$. The $w(\tilde{\mathbf{D}})$ (6.7) is minimized at $\tau = 4.9461 \times 10^{-9}$ with a value of $w(\tilde{\mathbf{D}}) = 41.1328$, and as $\tau$ increases the variance estimators $\tilde{v}(\tilde{\mathbf{D}})$ (6.9) and $\hat{v}(\tilde{\mathbf{D}})$ (6.10) diverge.

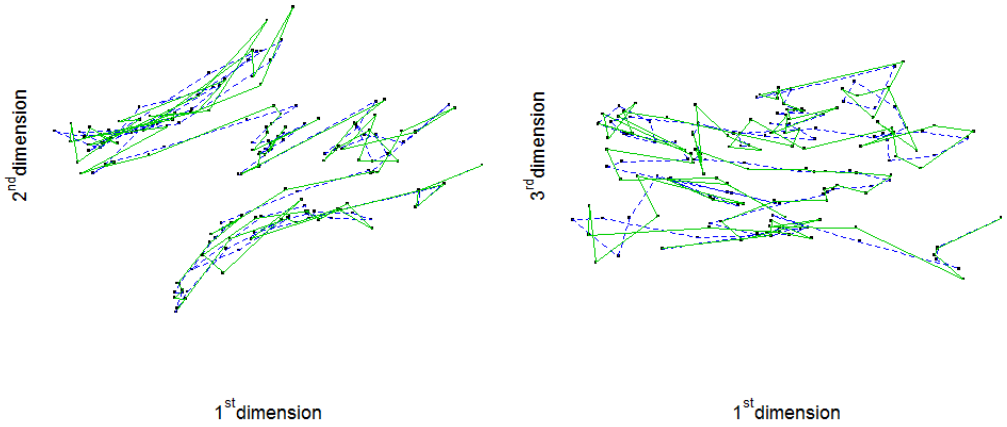
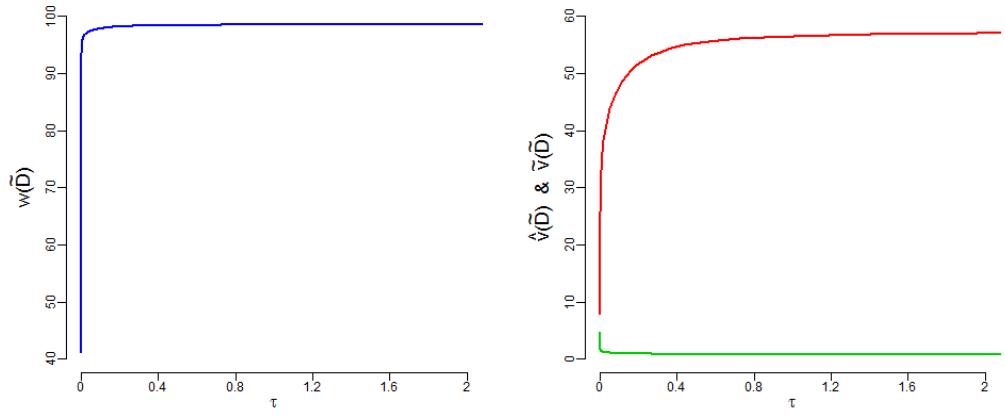



Figure 6.14: Perspectives of Chromosome 14's bias corrected and smoothed estimated configuration $\hat{\hat{\mathbf{X}}}^*_{P,M}$. The configuration $\hat{\hat{\mathbf{X}}}^*_{P,M}$ (Section 6.4.5) is found by applying the smoothing algorithm (Section 6.4.2) for the power transform (4.6), to Chromosome 14' bias corrected estimated configuration $\hat{\mathbf{X}}^*_{P,M}$. The configuration $\hat{\mathbf{X}}^*_{P,M}$ is found in Section 5.7. The point —●— denotes a megabase interval of $\hat{\hat{\mathbf{X}}}^*_{P,M}$ and the blue line connects successive megabase intervals of $\hat{\hat{\mathbf{X}}}^*_{P,M}$. The point -●- denotes a point of $\hat{\mathbf{X}}^*_{P,M}$ and the dashed green line connects successive points of $\hat{\mathbf{X}}^*_{P,M}$.




(a) $w(\tilde{\mathbf{D}})$. (b) $\tilde{v}(\tilde{\mathbf{D}}^*)$ and $\hat{v}(\tilde{\mathbf{D}}^*)$.

Figure 6.15: Chromosome 14's bias corrected and smoothed estimated configurations $\hat{\tilde{\mathbf{X}}}^*_{P,M}$ (Section 6.4.5) variance score values $w(\tilde{\mathbf{D}})$ (6.7), and variance estimates $\tilde{v}(\tilde{\mathbf{D}})$ (6.9) and $\hat{v}(\tilde{\mathbf{D}})$ (6.10). The variance score has a minimized value of $w(\tilde{\mathbf{D}}) = 41.1328$ at $\tau = 4.9461 \times 10^{-9}$. Left panel: the variance score $w(\tilde{\mathbf{D}})$ denoted by the blue line ___ as the smoothing parameter $\tau$ increases. Right panel: variance estimates $\tilde{v}(\tilde{\mathbf{D}})$ (6.9) denoted by the red line ___, and $\hat{v}(\tilde{\mathbf{D}})$ (6.10) denoted by the green line ___ as the smoothing parameter $\tau$ increases.

### 6.4.6 Conclusion

The post-processing through the smoothing algorithm proved unsuccessful when applied to $\hat{\mathbf{X}}$ from the exponential transform, and successful when applied to $\hat{\mathbf{X}}^*$ from the power transform. The smoothing algorithm failed for the exponential transform as the smooth algorithm over-smoothed $\hat{\mathbf{X}}$ producing $\hat{\tilde{\mathbf{X}}}$'s more dissimilar from $\mathbf{X}$. The $w(\tilde{\mathbf{D}})$ (6.7) for the exponential transform did not help in indicating where the smoothing had failed and held little relationship with the coefficient of variation. The smoothing algorithm was successful for the power transform, with only a small amount of smoothing but enough to improve $P(\mathbf{X}, \hat{\mathbf{X}})$. The $w(\tilde{\mathbf{D}})$ values from the smoothing algorithm for the power transform have a relationship with the coefficient of variation, providing some evidence of success. When the smoothing algorithm was applied to $\hat{\mathbf{X}}^*_{P,M}$ only minor adjustments were made leaving little shape difference between $\hat{\tilde{\mathbf{X}}}^*_{P,M}$ and $\hat{\mathbf{X}}^*_{P,M}$. The $w(\tilde{\mathbf{D}})$ was

minimized at $\tau = 4.9461 \times 10^{-9}$ suggesting improvements could still be made to the smoothing algorithm.

To improve the smoothing algorithm the variance score could be replaced with the absolute difference between $\tilde{v}(\tilde{\mathbf{D}})$ (6.5) and $\hat{v}(\tilde{\mathbf{D}})$ (6.6), such that

$$w(\tilde{\mathbf{D}}) = |\tilde{v}(\tilde{\mathbf{D}}) - \hat{v}(\tilde{\mathbf{D}})|.$$

Alternatively the smoothing algorithm could use a variance score based on the simple stress function $R_p(\hat{\mathbf{X}})$ (5.24) similar to when estimating the dispersion (in Section 5.5.1).

## 6.5 Score function investigation

In Chapter 4 the observed Hi-C counts (Lieberman-Aiden et al., 2009) were transformed into estimated distances, using a transform function with parameters found using the fitting algorithm (Section 4.1.3) with a specific score function. The estimated distances were then fitted into three dimensional Euclidean space using multidimensional scaling (MDS), to give an estimated chromosome configuration. The estimated chromosome configuration being the configuration which minimized the score function. When using non-metric MDS, the ordering of the fitted distances (2.2) from the estimated chromosome configuration were compared with the estimated distances through the stress of fit $S_p(\hat{\mathbf{X}})$ (2.14) score function. When using metric MDS the fitted counts recovered from the fitted distances using the inverse transforms (4.5) or (4.7), were compared with the observed Hi-C counts using the sum of the Pearsons residuals $\chi^2$ (4.9) score function. It was observed that changes in the small fitted distances become magnified to produce very large counts, and this could be biasing the $\chi^2$ values. To remedy this potential bias from the very large counts a distance based score function such as the stress $S_p(\hat{\mathbf{X}})$ (2.14) could be used for metric MDS. If $S_p(\hat{\mathbf{X}})$ was used as a score function for metric MDS the fitting

algorithm could seek an estimated configuration similar to the non-metric MDS solution, so the simple stress $R_p(\hat{\mathbf{X}})$ (5.24) can be used instead.

### 6.5.1 Transform function parameter estimation

To investigate which score function is preferable at finding the original transform parameters when using metric MDS a small set of exploratory simulations were run. These simulations involved generating a perturbed count matrix $\mathbf{M}$ using the procedure from the model based approach MBA (Chapter 5), with either the exponential transform (4.4) or power transform (4.6) and a level of dispersion $\rho$. Then using the fitting algorithm (Section 4.1.3) with metric MDS and one of the score functions $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24), to estimate $\hat{\alpha}$ for (4.4) or $\hat{\beta}$ for (4.6) and obtain a fitted configuration $\hat{\mathbf{X}}$. The fitted configuration $\hat{\mathbf{X}}$ can then be compared with the original configuration $\mathbf{X}$ using the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5). When using the power transform the bias correction can be applied using $\hat{\beta}$ in the estimate for the coefficients of inflation $\hat{\mathbf{C}} = (\hat{c}_{i,j})$ (5.26), to find if the bias correction is still applicable when using an estimated $\hat{\beta}$. The true value for dispersion is used in the estimation of (5.26) to provide more accuracy.

The original configuration $\mathbf{X}$ of a semi-circle is used in the simulations. When generating $\mathbf{M}$ using the exponential transform simulations were run to estimate $\alpha$ for $\alpha = 0.1, 0.01, 0.001$ and 0.0001. When using the power transform simulations were run to estimated $\beta$ for $\beta = -0.3, -0.5$ and $-0.7$ for each level of $b_0 = 0.1, 0.01, 0.001$ and 0.0001. The power transform simulations will use the minimum count adjustment of $m_{\text{min}} = 2$. When using the exponential transform $P(\mathbf{X}, \hat{\mathbf{X}})$ is measured between the original configuration $\mathbf{X}$ and the fitted configuration $\hat{\mathbf{X}}$. When using the power transform $P(\mathbf{X}, \hat{\mathbf{X}})$ is measured between the original configuration $\mathbf{X}$ and the fitted configuration $\hat{\mathbf{X}}$, and also measured between $\mathbf{X}$ and the bias corrected configuration $\hat{\mathbf{X}}^*$.

**Parameter estimation simulation results**

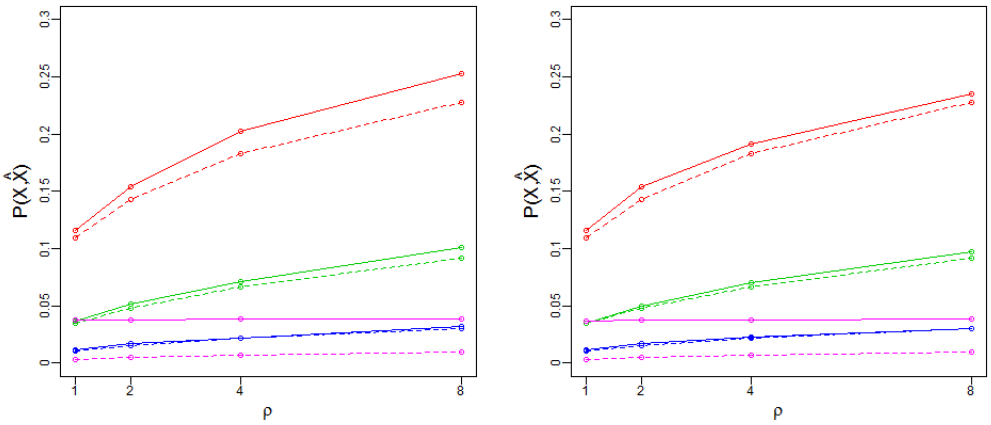
The results for the parameter estimation using $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score functions in the fitting algorithm, are displayed in Tables 6.3 and 6.4 and in Appendix Section F.

| $\rho = 1$ | $\chi^2$ | | $R_p(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $\alpha$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ |
| 0.1 | 0.087751 | 12.249 | 0.088738 | 11.2622 |
| 0.01 | 0.009869 | 1.3094 | 0.009828 | 1.7231 |
| 0.001 | 0.000999 | 0.1403 | 0.00099 | 1.0458 |
| 0.0001 | 0.000091 | 9.2949 | 0.00009 | 9.8301 |

Table 6.3: $\alpha$ estimates $(\hat{\alpha})$ from the parameter estimation simulations for the exponential transform (4.4). The $\hat{\alpha}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle, using the exponential transform and with dispersion $\rho = 1$. Column one gives the different levels of $\alpha$ used. For each score function the mean $\hat{\alpha}$ and the mean percentage error between $\hat{\alpha}$ and $\alpha$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_p(\hat{\mathbf{X}})$ score function.

When estimating $\alpha$ for the exponential transform the performance of the score function appears to depend on the size of $\alpha$. When $\alpha$ is a large size at $\alpha = 0.1$ the simple stress $R_p(\hat{\mathbf{X}})$ performs betters in the fitting algorithm, with the margin in performance between the score functions increasing as dispersion increases. When $\alpha$ is middling size at $\alpha = 0.01$ or 0.001 both score functions perform well in the fitting algorithm in estimating $\alpha$. For large and middling $\alpha$ the accuracy in the estimates improves as $\alpha$ decreases due to the decrease in the coefficient of variation $\mathrm{C}_v(\mu_{i,j}, \rho)$ (5.4) in the counts. When $\alpha$ is a small size at $\alpha = 0.0001$ the accuracy of the estimation deteriorates for both score functions $\chi^2$ and $R_p(\hat{\mathbf{X}})$. This deterioration could be due to the small $\alpha$ altering the count to distance relationship, such that biases appear when fitting large distances of counts. On each level of $\alpha$ used the accuracy of the estimation deteriorates when using either score function $\chi^2$ or $R_p(\hat{\mathbf{X}})$, as the dispersion increases.

Figure 6.16 gives the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values for the exponential transform (4.4) between $\mathbf{X}$ and $\hat{\mathbf{X}}$ found using $\hat{\alpha}$. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values found using $\hat{\alpha}$ are larger than the original $P(\mathbf{X}, \hat{\mathbf{X}})$ values for each level of $\alpha$. The decrease in estimation accuracy at $\alpha = 0.0001$ is reflected in $P(\mathbf{X}, \hat{\mathbf{X}})$ for both score functions $\chi^2$ and $R_p(\hat{\mathbf{X}})$, with the $P(\mathbf{X}, \hat{\mathbf{X}})$ for $\alpha = 0.0001$ poorer than $P(\mathbf{X}, \hat{\mathbf{X}})$ for $\alpha = 0.001$.




(a) $\chi^2$. (b) $R_p(\hat{\mathbf{X}})$.

Figure 6.16: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the parameter estimation simulations for the exponential transform (4.4). The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original $\mathbf{X}$ and fitted $\hat{\mathbf{X}}$ configurations. The fitted configuration $\hat{\mathbf{X}}$ is fitted from the perturbed count matrix $\mathbf{M}$, using $\hat{\alpha}$ in the exponential transform and metric MDS. The $\hat{\alpha}$ is found using the fitting algorithm with either $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score functions. The matrix $\mathbf{M}$ is generated using the MBA approach, from a semi-circle using the original $\alpha$ values in (4.5). Left panel: $\chi^2$ used in the fitting algorithm. Right panel: $R_p(\hat{\mathbf{X}})$ used in the fitting algorithm. The red lines —○— for $\alpha = 0.1$; the green lines —○— for $\alpha = 0.01$; the blue lines —○— for $\alpha = 0.001$, and the pink lines —○— for $\alpha = 0.0001$. The dashed lines -○- give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3a.

The power transform uses two parameters $b_0$ and $\beta$. The parameter $b_0$ is used in the MBA to determine the size of the counts and therefore control the coefficient of variation $\mathrm{C}_v(\mu_{i,j}, \rho)$ in the perturbed counts. The parameter $\beta$ determines the shape of the relationship between counts and distances and this parameter requires accurate estimation for the power transform to be successful. The simulations were run to estimate $\beta$ at

$\beta = -0.3, -0.5$ and $-0.7$, to find if accuracy of the estimate $\hat{\beta}$ changed as $\beta$ changed.

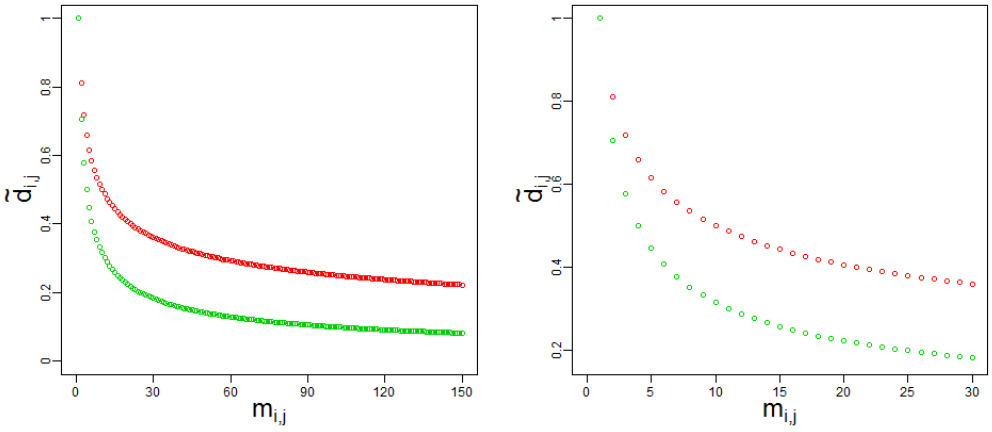
| $\rho = 1$ | $\chi^2$ | | $R_p(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.413679 | 17.2642 | -0.415072 | 16.9857 |
| 0.01 | -0.483658 | 3.2684 | -0.48861 | 2.2781 |
| 0.001 | -0.500732 | 0.1463 | -0.498822 | 0.2355 |
| 0.0001 | -0.499966 | 0.0069 | -0.499903 | 0.0194 |

Table 6.4: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.5$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 1$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_p(\hat{\mathbf{X}})$ score function.

When estimating $\beta$ both score functions performed well, with accuracy improving as the $\mathrm{C}_v(\mu_{i,j}, \rho)$ decreased (when $b_0$ and $\rho$ decreased). The most interesting results are found at the $b_0 = 0.1$ level as the $\mathrm{C}_v(\mu_{i,j}, \rho)$ is large, where the effect of the bias on $P(\mathbf{X}, \hat{\mathbf{X}})$ is also large. Therefore additional attention will be given to the performance of the score functions $\chi^2$ and $R_p(\hat{\mathbf{X}})$ in estimating $\beta$ at the $b_0 = 0.1$ level. When $\beta = -0.3$ the most accurate $\beta$ estimates come from $R_p(\hat{\mathbf{X}})$ when dispersion is small $\rho \leq 2$ and then from $\chi^2$ when dispersion is large $\rho > 2$. When $\beta = -0.5$ the most accurate $\beta$ estimates come from $R_p(\hat{\mathbf{X}})$ when dispersion is small $\rho = 1$ and then from $\chi^2$ when dispersion is larger $\rho \geq 2$. When $\beta = -0.7$ the most accurate $\beta$ estimates come from $\chi^2$ on all levels of dispersion.

In all the simulations run using the $R_p(\hat{\mathbf{X}})$ score function and most of the simulations run using the $\chi^2$ score function, the $\hat{\beta}$ values are greater than the true $\beta$ values. The difference between the $\hat{\beta}$ and $\beta$ values continues to increase, as $\beta$ decreases, this can be observed in the increase in the mean percentage error values for the estimates. Transforming counts

into distances using $\hat{\beta} > \beta$ (with $b_0 = 1$) using the power transform (Figure 6.17) has two effects on the resulting distances. One effect is the distances generated using $\hat{\beta}$ are larger than the distances generated using $\beta$. The other effect is the size difference between the distances generated using $\hat{\beta}$ is smaller than the size difference between distances generated using $\beta$. This second effect has an impact on distances generated from small counts where the coefficient of variation is larger, as it reduces the size of the noise around large distances.




(a) (b)

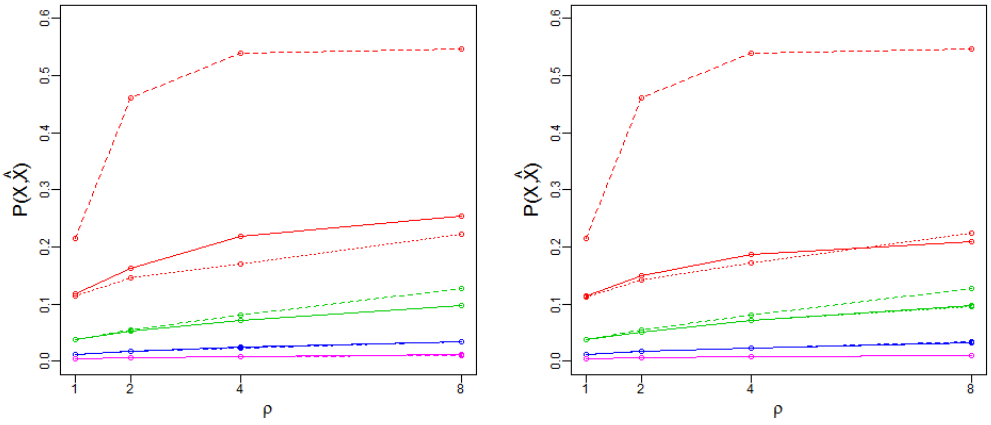
Figure 6.17: Illustration of how distances change when using a greater value of $\beta$ in the power transform (4.6). The red circles ○ denote distances transformed from counts using $\beta = -0.3$ and $b_0 = 1$ in the power transform. The green circles ○ denote distances transformed from counts using $\beta = -0.5$ and $b_0 = 1$ in the power transform.

Figure 6.18 gives the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ values for the power transform with $\beta = -0.5$, between $\mathbf{X}$ and $\hat{\mathbf{X}}$ and between $\mathbf{X}$ and $\hat{\mathbf{X}}^*$ found using $\hat{\beta}$. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values found when using $\hat{\beta}$ from $\chi^2$ are much smaller than the original $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA (in Appendix Section B), and at the larger levels of $\mathrm{C}_v(\mu_{i,j}, \rho)$ are smaller than the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA bias corrected configuration $\hat{\mathbf{X}}^*$ (in Appendix Section D). Applying the bias correction to the perturbed distances using the $\hat{\beta}$ estimated using $\chi^2$ still manages to provide an improve $P(\mathbf{X}, \hat{\mathbf{X}})$ when the $\mathrm{C}_v(\mu_{i,j}, \rho)$ is large, and provides a negligible difference when the $\mathrm{C}_v(\mu_{i,j}, \rho)$ is small. Therefore using a $\hat{\beta}$ found with $\chi^2$

to recover $\hat{\mathbf{X}}$ proves more beneficial than using the original $\beta$ and the bias correction. Using a $\hat{\beta}$ found with $\chi^2$ in the bias correction to recover $\hat{\mathbf{X}}^*$, still provides improvement in the recovery of the original configuration from the perturbed count matrix $\mathbf{M}$. The shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ values found when using $\hat{\beta}$ from $R_p(\hat{\mathbf{X}})$ are much smaller than the original $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA, and at large $\mathrm{C}_v(\mu_{i,j}, \rho)$ are smaller than MBA bias corrected configuration $P(\mathbf{X}, \hat{\mathbf{X}})$ values. Applying the bias correction to the perturbed distances provides improvement at large $\mathrm{C}_v(\mu_{i,j}, \rho)$ although fails when $b_0 = 0.1$ and $\rho = 8$, and at smaller $\mathrm{C}_v(\mu_{i,j}, \rho)$ the difference in $P(\mathbf{X}, \hat{\mathbf{X}})$ is negligible. Therefore using a $\hat{\beta}$ found with $R_p(\hat{\mathbf{X}})$ to recover $\hat{\mathbf{X}}$ proves more beneficial than using the original $\beta$ and the bias correction. Using a $\hat{\beta}$ found with $R_p(\hat{\mathbf{X}})$ in the bias correction to recover $\hat{\mathbf{X}}^*$, still provides improvement in the recovery of the original configuration from the perturbed count matrix $\mathbf{M}$, most of the time.

### 6.5.2 Crossing the transform functions

In the model based approach (MBA) perturbed distances $\tilde{\mathbf{D}} = (\tilde{d}_{i,j})$ are generated by taking a configurations true distances $\mathbf{D} = (d_{i,j})$ and transforming them into mean counts $\mathbf{U} = (\mu_{i,j})$ using an inverse exponential (4.5) or inverse power (4.7) transform function. The $\mathbf{U}$ are then used to simulate perturbed counts $\mathbf{M} = (m_{i,j})$ using the Poisson or negative binomial distribution. The $\mathbf{M}$ are transformed into $\tilde{\mathbf{D}}$ using the same transform function, either the exponential transform (4.4) or power transform (4.6). For example if the power transform inverse (4.7) was used to obtain $\mathbf{U}$ then the power transform (4.6) is used to obtain $\tilde{\mathbf{D}}$, with the same symmetry when using the exponential transform. This section will investigate the effects of crossing the transform functions; by producing $\mathbf{U}$ with either the inverse exponential transform or the inverse power transform, then producing $\tilde{\mathbf{D}}$ with the other transform function, either the power transform or the exponential transform. This is relevant when transforming the observed Hi-C counts (Lieberman-Aiden et al., 2009) as it is assumed the correct transform is used through




(a) $\chi^2$. (b) $R_p(\hat{\mathbf{X}})$.

Figure 6.18: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the parameter estimation simulations for the power transform (4.6). The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original $\mathbf{X}$ and fitted $\hat{\mathbf{X}}$ configurations. The fitted configuration $\hat{\mathbf{X}}$ is fitted from the perturbed count matrix $\mathbf{M}$, using $\hat{\beta}$ in the power transform; with the $m_{\text{min}} = 2$ adjustment and metric MDS. The $\hat{\beta}$ is found using the fitting algorithm with either $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score functions. The matrix $\mathbf{M}$ is generated using the MBA approach, from a semi-circle using the original $\beta = -0.5$ value in (4.7). Left panel: $\chi^2$ used in the fitting algorithm. Right panel: $R_p(\hat{\mathbf{X}})$ used in the fitting algorithm. The red lines for $b_0 = 0.1$; the green lines for $b_0 = 0.01$; the blue lines for $b_0 = 0.001$, and the pink lines for $b_0 = 0.0001$. The dotted red line gives the $P(\mathbf{X}, \hat{\mathbf{X}})$ value between $\mathbf{X}$ and the bias corrected $\hat{\mathbf{X}}^*$ using $\hat{\beta}$ for $b_0 = 0.1$. The dashed lines give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3b.

either the exponential transform or the power transform. Here a small exploratory set of simulations are run, which generate $\mathbf{M}$ using one transform inverse and transforming back to $\tilde{\mathbf{D}}$ using the other transform. Then fitting $\tilde{\mathbf{D}}$ into Euclidean space using metric multidimensional scaling (MDS) to obtain a fitted configuration $\hat{\mathbf{X}}$. The shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) between the original configuration $\mathbf{X}$ and $\hat{\mathbf{X}}$ will be measured, and compared with the $P(\mathbf{X}, \hat{\mathbf{X}})$ from the MBA simulations.

The original configuration $\mathbf{X}$ of a semi-circle will be used in the simulations. When using the inverse exponential transform simulations will be run at different levels of $\alpha$ at $\alpha = 0.1, 0.01, 0.001$ and 0.0001; the perturbed counts $\mathbf{M}$ will be transformed into $\tilde{\mathbf{D}}$ using the power transform with $m_{\text{min}} = 2$, $b_0 = 0.1$ and $\beta$ estimated using the fitting algorithm with either $\chi^2$ or $R_p(\hat{\mathbf{X}})$ score function. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values form the simulation using the exponential transform inverse, are compared with the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA simulations using the exponential transform (Figure 5.3a). The bias correction is omitted as the nature of the counts is different. When using the power transform inverse simulations will be run at different levels of $b_0$ at $b_0 = 0.1, 0.01, 0.001$ and 0.0001, $\beta = -0.5$; the perturbed counts will be transformed into $\tilde{\mathbf{D}}$ using the exponential transform with $\alpha$ estimated using the fitting algorithm with either $\chi^2$ or $R_p(\hat{\mathbf{X}})$ score function. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the simulations using the power transform inverse, are compared with the $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA simulations using the power transform (Figure 5.3b). The simulations will be run for each level of dispersion $\rho = 1, 2, 4$ and 8. Figure 6.19 gives the $P(\mathbf{X}, \hat{\mathbf{X}})$ plots when using the inverse exponential transform (4.5) and the power transform (4.6) to obtain $\hat{\mathbf{X}}$. When using the $\chi^2$ (4.9) in the fitting algorithm to estimate $\beta$ the $P(\mathbf{X}, \hat{\mathbf{X}})$ values are ordered opposite to the level of $\text{C}_v(\mu_{i,j}, \rho)$ (5.4) in the perturbed counts, so $P(\mathbf{X}, \hat{\mathbf{X}})$ deteriorates when $\alpha$ decreases (and as $\text{C}_v(\mu_{i,j}, \rho)$ decrease). At the $\alpha = 0.1$ level the $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values almost match the original $P(\mathbf{X}, \hat{\mathbf{X}})$ values from the MBA. At the $\alpha = 0.1$ level the fitting algorithm and $\chi^2$ score function estimated $\hat{\beta} = -0.8843$ (when dispersion $\rho = 1$). The plot of the relationship between counts and distances the exponential transform makes at $\alpha = 0.1$, and the

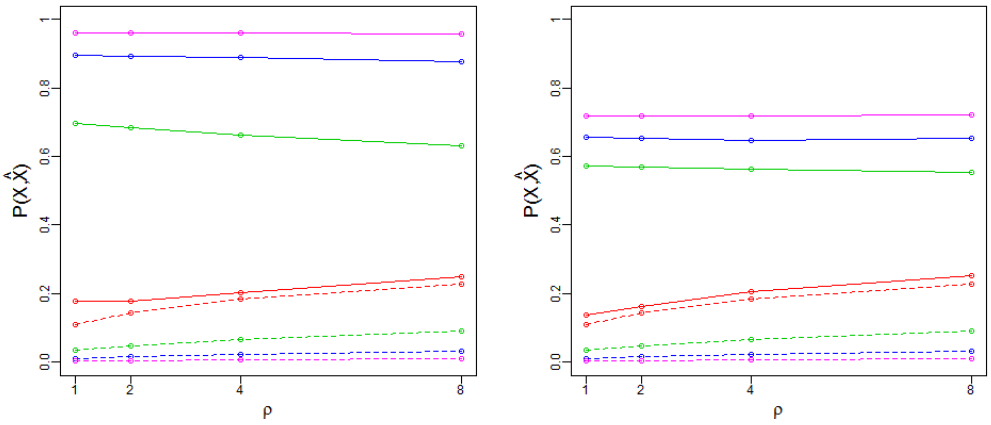
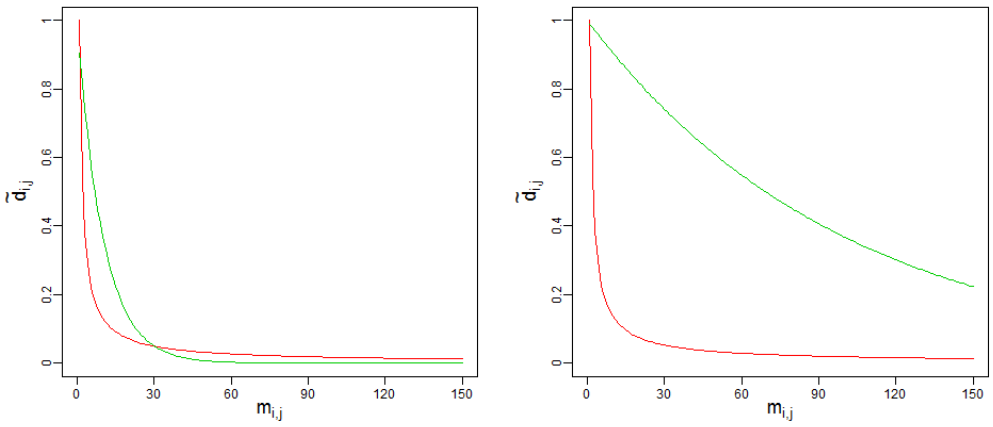


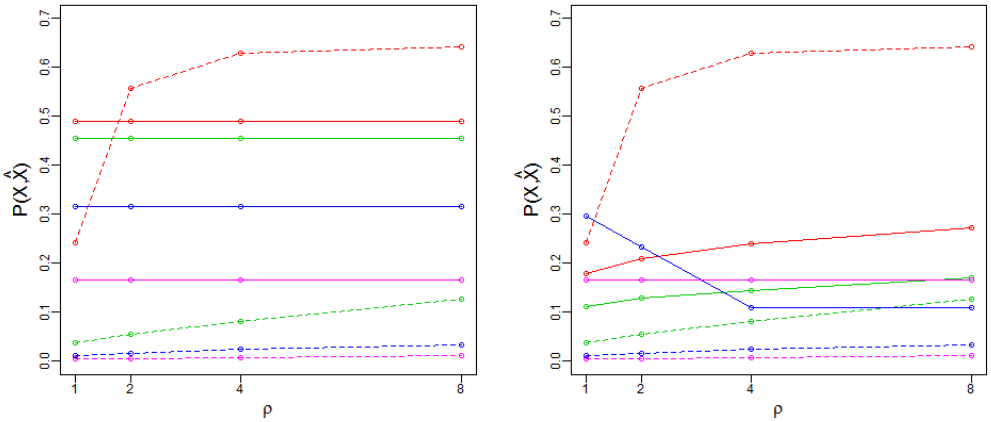

Figure 6.19: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the parameter estimation simulations for crossing transforms (exponential transform to power transform). The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original $\mathbf{X}$ and fitted $\hat{\mathbf{X}}$ configurations. The fitted configuration $\hat{\mathbf{X}}$ is fitted from the perturbed count matrix $\mathbf{M}$, with the $m_{\text{min}} = 2$ adjustment; using $\hat{\beta}$ in the power transform (4.6) and metric MDS. The $\hat{\beta}$ is found using the fitting algorithm with either $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score functions. The matrix $\mathbf{M}$ is generated using the MBA approach, from a semi-circle using the original $\alpha$ value in (4.5). Left panel: $\chi^2$ used in the fitting algorithm. Right panel: $R_p(\hat{\mathbf{X}})$ used in the fitting algorithm. The red lines —○— for $\alpha = 0.1$; the green lines —○— for $\alpha = 0.01$; the blue lines —○— for $\alpha = 0.001$, and the pink lines —○— for $\alpha = 0.0001$. The dashed lines - ○ - give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3a.

relationship between counts and distances the power transform makes at $\hat{\beta} = -0.8843$, is displayed in Figure 6.20a. The relationship both transforms make between counts and distances looks similar. The close $P(\mathbf{X}, \hat{\mathbf{X}})$ values will be due ability of the power transform to capture the relationship the exponential transform makes between counts and distances at $\alpha = 0.1$. At the other levels of $\alpha$ ($= 0.01, 0.001$ and $0.0001$) the relationship between the counts and distances is captured poorly by the power transform. This poor fit can be observed in Figure 6.20b, where the counts are generated using the inverse of the exponential transform (4.5); perturbed with dispersion set at $\rho = 1$ and returned to perturbed distances using the power transform (4.6) with $\beta = -0.7207$. When using the $R_p(\hat{\mathbf{X}})$ (5.24) in the fitting algorithm to estimate $\hat{\beta}$ the $P(\mathbf{X}, \hat{\mathbf{X}})$ values are similar to when $\chi^2$ is used in the fitting algorithm, but with a marginal improvement in $P(\mathbf{X}, \hat{\mathbf{X}})$. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values deteriorate as $\mathrm{C}_v(\mu_{i,j}, \rho)$ decreases when $\alpha$ decreases, and the $P(\mathbf{X}, \hat{\mathbf{X}})$ values at $\alpha = 0.1$ are a close match to the original MBA $P(\mathbf{X}, \hat{\mathbf{X}})$ values.




(a) Good emulation. (b) Poor emulation.

Figure 6.20: Illustration of how the relationships of the transforms emulate each other. Left panel: the relationship between counts and distances from using the exponential transform (4.4) with $\alpha = 0.1$ and using the power transform (4.6) with $\beta = -0.8843$. Right panel: the relationship between counts and distances from using the exponential transform with $\alpha = 0.01$ and using the power transform with $\beta = -0.7207$. The red line — denotes distances generated from the exponential transform. The green line — denotes distances generated from the power transform.




(a) $\chi^2$. (b) $R_p(\hat{\mathbf{X}})$.

Figure 6.21: Shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) values from the parameter estimation simulations for crossing transforms (power transform to exponential transform). The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original $\mathbf{X}$ and fitted $\hat{\mathbf{X}}$ configurations. The fitted configuration $\hat{\mathbf{X}}$ is fitted from the perturbed count matrix $\mathbf{M}$, using $\hat{\alpha}$ in the exponential transform (4.4) and metric MDS. The $\hat{\alpha}$ is found using the fitting algorithm with either $\chi^2$ (4.9) or $R_p(\hat{\mathbf{X}})$ (5.24) score functions. The matrix $\mathbf{M}$ is generated using the MBA approach, from a semi-circle using the original $\beta = -0.5$ value in the inverse power transform (4.7). Left panel: $\chi^2$ used in the fitting algorithm. Right panel: $R_p(\hat{\mathbf{X}})$ used in the fitting algorithm. The red lines —○— for $b_0 = 0.1$; the green lines —○— for $b_0 = 0.01$; the blue lines —○— for $b_0 = 0.001$, and the pink lines —○— for $b_0 = 0.0001$. The dashed lines - ○ - give the equivalent MBA simulation $P(\mathbf{X}, \hat{\mathbf{X}})$ values from Figure 5.3b.

Figure 6.21 gives the $P(\mathbf{X}, \hat{\mathbf{X}})$ plots when using the inverse of the power transform (4.7) and the exponential transform (4.4) to obtain $\hat{\mathbf{X}}$. When using the $\chi^2$ in the fitting algorithm to estimate $\alpha$ the $P(\mathbf{X}, \hat{\mathbf{X}})$ values are poorer than the original MBA $P(\mathbf{X}, \hat{\mathbf{X}})$ values, this the exception of $P(\mathbf{X}, \hat{\mathbf{X}})$ for $b_0 = 0.1$. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values improve as $\mathrm{C}_v(\mu_{i,j}, \rho)$ decreases when $b_0$ decreases and appear completely unaffected by the level of dispersion. When using $R_p(\hat{\mathbf{X}})$ in the fitting algorithm to estimate $\alpha$, the $P(\mathbf{X}, \hat{\mathbf{X}})$ values are an improvement to when $\chi^2$ is used in the fitting algorithm, with the $P(\mathbf{X}, \hat{\mathbf{X}})$ values deteriorating as the dispersion increases. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values at $b_0 = 0.1$ appears much lower than the original MBA $P(\mathbf{X}, \hat{\mathbf{X}})$ values.

### 6.5.3 Conclusion

Estimating $\alpha$ or $\beta$ using a count based score function $\chi^2$ (4.9) or a distance based score function $R_p(\hat{\mathbf{X}})$ (5.24) has provided additional insight useful in fitting the Hi-C count matrices.

When estimating $\alpha$ for the exponential transform (4.4) both score functions provide accurate estimates. Although the accuracy in the estimates deteriorates at very small $\alpha$ (at $\alpha = 0.0001$). The effect of using $\hat{\alpha}$ in producing the fitted configuration is a deterioration in the shape difference $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5), with the size of the deterioration reflected in the accuracy of $\hat{\alpha}$ at estimating $\alpha$.

When estimating $\beta$ for the power transform (4.6) both score functions perform well, with the accuracy of the estimate neatly linked to the coefficient of variation $\mathrm{C}_v(\mu_{i,j}, \rho)$ (5.4) in the counts (improving as $\mathrm{C}_v(\mu_{i,j}, \rho)$ decreased). The use of either score function in the fitting algorithm produced $\hat{\beta}$ values greater than the original $\beta$ values. Using a greater $\hat{\beta}$ instead of $\beta$ reduces the amount of variation in the small distances which improves the recovery of the fitted configuration. Using $\hat{\beta}$ in estimating the coefficient of inflation (5.26) to use to give a bias corrected configuration $\hat{\mathbf{X}}^*$, still improved $P(\mathbf{X}, \hat{\mathbf{X}})$.

The bias correction was not successful when $\hat{\beta}$ was estimated using $R_p(\hat{\mathbf{X}})$ in the score function at $b_0 = 0.1$ and $\rho = 8$, although at this level it was show using $\chi^2$ provided the more accurate $\hat{\beta}$. The $\hat{\beta}$ used when fitting the observed Hi-C count matrix (Lieberman-Aiden et al., 2009) for Chromosome 14 could be much greater than the hypothetical $\beta$, where the hypothetical $\beta$ is used to transform the distance between the megabase intervals into the hypothetical mean counts.

The use of the same transform parameter used in the generation of $\mathbf{M}$, might not be ideal in the recovery of a fitted configuration which resembles the original configuration. More broader simulations including the use of the straight-line; parabola and circle could be carried out to see if the $\hat{\beta} > \beta$ and if the bias correction still works when using $\hat{\beta}$. The properties of the perturbed distances when using an incorrect transform function parameter $\hat{\alpha} \neq \alpha$ (for the exponential transform (4.4)) or $\hat{\beta} \neq \beta$ (for the power transform (4.6)), could be investigated further through the delta-method. This could improve the bias correction technique or find a bias correction which avoids having to pre-process the perturbed distances.

Crossing the transform functions provides insight into the effects of using the wrong transform function to obtain the estimated distances.

When using the power transform to emulate the exponential transform, the $P(\mathbf{X}, \hat{\mathbf{X}})$ values deteriorate as the $\mathrm{C}_v(\mu_{i,j}, \rho)$ decreases. This deterioration could be linked to the inability of the power transform to emulate the exponential transform (illustrated in Figure 6.20).

When using the exponential transform to emulate the power transform, the $P(\mathbf{X}, \hat{\mathbf{X}})$ values at large $\mathrm{C}_v(\mu_{i,j}, \rho)$ when $b_0 = 0.1$ are an improvement on the original MBA $P(\mathbf{X}, \hat{\mathbf{X}})$ values. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values for the other levels of $b_0$ at $b_0 = 0.01, 0.001$ and 0.0001 are a deterioration on the original MBA $P(\mathbf{X}, \hat{\mathbf{X}})$ values.

Using the wrong transform function can lead to a deterioration in the $P(\mathbf{X}, \hat{\mathbf{X}})$ values as

the transform function cannot properly capture the original count to distance relationship. One approach to further studying using the wrong transform function would be to produce distorted distances, by obtaining the mean counts through either the exponential or power transform inverse and transform back with the power or exponential transform. The distorted distances would avoid the perturbation stage and provide a starting point for investigating the properties of distances from using the wrong transform function. Using different more complex transform functions such as adapting the Box-Cox function, could provide transform functions which emulate the original count to distance relationship.

# Chapter 7

# Critical summary and directions for future research

## 7.1 Critical summary

This thesis applied multidimensional scaling (MDS) to Hi-C data from male human lymphoblast cells to recover and estimated chromosome (or genome) configuration. Two contact-count to distance transform functions were used with two methods of MDS providing four estimated chromosome configurations, three of which shared the same shape. One prominent feature in the estimated chromosome configurations was a horseshoe shape caused either by a genuine horseshoe structure of the chromosome, or most probably by the horseshoe effect. The estimated chromosome configurations were investigated to try detect known features of the chromosome. The chromosome contact matrix from Chromosome 14 displays a plateauing of contact-counts at medium to large genomic distances, making the estimation of Euclidean distances difficult on the plateau, providing an ingredient for the horseshoe effect. Estimates of genome configuration capture the established concept of chromosome territory, but fail to capture experimentally observed radial positioning of the chromosomes.

The mechanics of metric MDS were also investigated to glean information which could aid the recovery of the chromosome configuration and gain insight into the negative eigenvalues found when fitting non-Euclidean distance matrices.

Acknowledging that the true chromosome configuration is unknown, a model-based approach (MBA) was developed which simulated contact-counts data from predefined configurations. The MBA identified a bias in both transform functions and tried to gauge its affect in the fitted configurations. A bias-correction technique and dispersion estimation method was developed for the power transform when fitting with metric MDS and applied with success. The bias correction was then applied to the Hi-C data and produced a reduction in the score function used to measure fit and subtle alterations in local structure of the estimated chromosome configuration. The affect of the bias on the fitted configurations and eigenvalues was investigated; highlighting that the horseshoe in the estimated chromosome configuration found using the exponential transform could be partly driven by the bias. Combinations of proposed bias correction techniques for the exponential transform were trialled with no success, leading to the conclusion that bias cannot easily be corrected for when using the exponential transform. Smoothing splines were used to post-process the fitted configurations but only proved successful when using the power transform. An exploratory investigation into which score function performed better in the fitting algorithm to estimate the original transform parameters was undertaken. This found that when estimating $\alpha$ for the exponential transform both score functions performed well at providing a $\hat{\alpha}$ close to the original $\alpha$. The exploratory investigation also found that estimating $\beta$ for the power transform the $\hat{\beta}$ values were usually greater than the original $\beta$ values, although using $\hat{\beta}$ in the bias correction still proved to improve the fitted configuration.

## 7.2 Directions for future research

Refining the fitting procedure, modelling for the Hi-C data to represent a cell average of genome configurations and multiple resolution multidimensional scaling are three directions for future research.

Refining the fitting procedure first involves investigating the effects of using incorrect transform parameters and how this effects the bias correction for the power transform. Another refinement would be applying the procedure outlined in the trials Section 6.2 to find a better-performing bias correction, for the power transform when fitting into Euclidean space with metric MDS. Another refinement would be making the bias correction for the power transform recursive. This should improve the recovery of the fitted configuration from the perturbed counts, since after each bias correction improved fitted distances and counts should obtained, which can be used to give an improved coefficient of inflation and improve bias corrected perturbed (or estimated) distances. These improvements can be compounded to aid the recovery of the estimated chromosome configuration.

The Hi-C data represents a cell average of genome configurations so the estimated chromosome configuration is also an average of these configurations. Using the same approach as the MBA, multiple configurations can be used to produce an average count matrix and corresponding perturbed distance matrix, which can be fitted into Euclidean space using MDS, to give an average fitted configuration. The average fitted configuration can be compared with the original configurations, to find which it original configuration it resembles the most. Methods might then be developed to extract the multiple configurations from the average contact matrix which can then be applied to the Hi-C data.

Multiple resolution multidimensional scaling involves reducing the resolution of the chromosome contact-count matrix from Hi-C or the perturbed contact matrix from the

MBA, and recovering a fitted configuration from the new matrix. This would involve applying the procedure outlined in Section 4.8.1 for fitting the genome to the low-resolution contact-count matrix, taking the lower resolution contact-count matrix and extracting a low-resolution fitted configuration. The low-resolution fitted configuration can provide insight into how robust the transforming of contact-count to distances and MDS is to changes in resolution.

## 7.3 Conclusion

This thesis has shown that Hi-C data can be transformed into estimated distances, and an estimated chromosome configuration can be fitted from the estimated distances using multidimensional scaling. Although the estimated chromosome configurations found using this procedure are susceptible to the horseshoe effect, which is caused by properties of the transform function and the nature of the Hi-C data. Attempts were made to correct and gauge the transforms effect on the estimated chromosome configuration by using a model based approach as a platform for investigation. Therefore, in conclusion, sensible investigation of the transform functions and multidimensional scaling is required to minimize the extent of the horseshoe effect in the estimated chromosome configuration.

# Appendices

## A Chromosome score function data

Appendix containing the score function data when obtaining the estimated chromosome configurations with metric or non-metric multidimensional scaling (MDS), for all 22 chromosomes and the X chromosome.

### A.1 Chromosome score function data from metric MDS

Appendix containing the score function data when obtaining the estimated chromosome configurations using the exponential transform (4.4) or the power transform (4.6), and using metric MDS.

| Chromosome | Transform function | Parameter | Parameter estimate | $\chi^2$ | SSR$(\mathbf{M}, \hat{\mathbf{U}})$ | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|---|---|
| 1 | Exponential | $\hat{\alpha}$ | 0.0167 | 868908 | $1.086\times10^8$ | 17.3677 |
| | Power | $\hat{\beta}$ | -0.4982 | 1445467 | $2.4202\times10^9$ | 19.8895 |
| 2 | Exponential | $\hat{\alpha}$ | 0.0159 | 9441425 | $1.0712\times10^8$ | 18.1352 |
| | Power | $\hat{\beta}$ | -0.506 | 1542520 | $1.618\times10^9$ | 20.1427 |
| 3 | Exponential | $\hat{\alpha}$ | 0.0155 | 568775 | $7.0007\times10^7$ | 15.9176 |
| | Power | $\hat{\beta}$ | -0.5163 | 1188609 | $1.357\times10^9$ | 21.2583 |
| 4 | Exponential | $\hat{\alpha}$ | 0.0171 | 436020 | $5.5748\times10^7$ | 16.8751 |
| | Power | $\hat{\beta}$ | -0.5137 | 1135280 | $1.1494\times10^9$ | 20.8081 |
| 5 | Exponential | $\hat{\alpha}$ | 0.0147 | 509040 | $6.8704\times10^7$ | 15.3901 |
| | Power | $\hat{\beta}$ | -0.4971 | 1023446 | $1.8224\times10^9$ | 20.3237 |
| 6 | Exponential | $\hat{\alpha}$ | 0.0141 | 493443 | $6.645\times10^7$ | 16.0704 |
| | Power | $\hat{\beta}$ | -0.4914 | 1121189 | $3.8213\times10^9$ | 20.3977 |
| 7 | Exponential | $\hat{\alpha}$ | 0.0130 | 469453 | $6.2335\times10^7$ | 18.7016 |
| | Power | $\hat{\beta}$ | -0.5029 | 1243815 | $1.7385\times10^9$ | 24.8679 |

Table A.1: Score function data for Chromosomes 1 to 7's estimated configurations, found using metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $\chi^2$ (4.9) value it minimizes. Column six and seven give the SSR$(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) and $S_3(\hat{\mathbf{X}})$ (2.14) values found using the estimated parameter values. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome's Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 2$ adjustment, and fitting into three dimensional space with metric MDS. The statistics $\chi^2$, SSR$(\mathbf{M}, \hat{\mathbf{U}})$ and $S_3(\hat{\mathbf{X}})$ are then extracted from the fitted configuration.

| Chromosome | Transform function | Parameter | Parameter estimate | $\chi^2$ | SSR$(\mathbf{M}, \hat{\mathbf{U}})$ | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|---|---|
| 8 | Exponential | $\hat{\alpha}$ | 0.0123 | 362217 | $5.1637\times10^7$ | 15.6740 |
| | Power | $\hat{\beta}$ | -0.4938 | 744585 | $6.6296\times10^8$ | 23.1395 |
| 9 | Exponential | $\hat{\alpha}$ | 0.0116 | 276317 | $4.6751\times10^7$ | 17.1640 |
| | Power | $\hat{\beta}$ | -0.4864 | 1426884 | $3.9002\times10^9$ | 24.1512 |
| 10 | Exponential | $\hat{\alpha}$ | 0.0114 | 344447 | $5.625\times10^7$ | 16.6842 |
| | Power | $\hat{\beta}$ | -0.4929 | 765634 | $1.0881\times10^9$ | 22.0247 |
| 11 | Exponential | $\hat{\alpha}$ | 0.0125 | 287594 | $4.4483\times10^7$ | 16.0595 |
| | Power | $\hat{\beta}$ | -0.4556 | 1253663 | $1.3172\times10^{10}$ | 25.0642 |
| 12 | Exponential | $\hat{\alpha}$ | 0.0124 | 297651 | $4.3647\times10^7$ | 16.4353 |
| | Power | $\hat{\beta}$ | -0.4632 | 709648 | $1.9794\times10^9$ | 22.9011 |
| 13 | Exponential | $\hat{\alpha}$ | 0.011 | 171307 | $2.8074\times10^7$ | 16.8519 |
| | Power | $\hat{\beta}$ | -0.5903 | 738917 | $7.5705\times10^8$ | 31.7842 |
| 14 | Exponential | $\hat{\alpha}$ | 0.0095 | 147172 | $2.7044\times10^7$ | 16.8774 |
| | Power | $\hat{\beta}$ | -0.4796 | 485658 | $7.4257\times10^8$ | 23.9336 |
| 15 | Exponential | $\hat{\alpha}$ | 0.0088 | 143456 | $3.0985\times10^7$ | 17.9886 |
| | Power | $\hat{\beta}$ | -0.5044 | 893767 | $2.7319\times10^9$ | 28.4703 |

Table A.2: Score function data for Chromosomes 8 to 15's estimated configurations, found using metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $\chi^2$ (4.9) value it minimizes. Column six and seven give the SSR$(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) and $S_3(\hat{\mathbf{X}})$ (2.14) values found using the estimated parameter values. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome's Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 2$ adjustment, and fitting into three dimensional space with metric MDS. The statistics $\chi^2$, SSR$(\mathbf{M}, \hat{\mathbf{U}})$ and $S_3(\hat{\mathbf{X}})$ are then extracted from the fitted configuration.

| Chromosome | Transform function | Parameter | Parameter estimate | $\chi^2$ | SSR$(\mathbf{M}, \hat{\mathbf{U}})$ | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|---|---|
| 16 | Exponential | $\hat{\alpha}$ | 0.0072 | 182872 | $4.1018\times10^7$ | 16.6576 |
| | Power | $\hat{\beta}$ | -0.4008 | 850103 | $5.8201\times10^9$ | 24.9965 |
| 17 | Exponential | $\hat{\alpha}$ | 0.0084 | 157526 | $3.036\times10^7$ | 17.4508 |
| | Power | $\hat{\beta}$ | -0.4542 | 677167 | $4.5271\times10^9$ | 27.0713 |
| 18 | Exponential | $\hat{\alpha}$ | 0.0088 | 112784 | $2.18\times10^7$ | 18.534 |
| | Power | $\hat{\beta}$ | -0.5667 | 424011 | $4.7909\times10^8$ | 31.6188 |
| 19 | Exponential | $\hat{\alpha}$ | 0.0065 | 70140 | $1.7563\times10^7$ | 13.8996 |
| | Power | $\hat{\beta}$ | -0.4682 | 781721 | $1.9125\times10^9$ | 33.8213 |
| 20 | Exponential | $\hat{\alpha}$ | 0.0064 | 79915 | $2.1531\times10^7$ | 14.8062 |
| | Power | $\hat{\beta}$ | -0.5076 | 581958 | $9.7289\times10^8$ | 32.9016 |
| 21 | Exponential | $\hat{\alpha}$ | 0.0051 | 18880 | $5.7184\times10^6$ | 13.7343 |
| | Power | $\hat{\beta}$ | -0.4116 | 162856 | $2.4729\times10^8$ | 21.098 |
| 22 | Exponential | $\hat{\alpha}$ | 0.0024 | 30214 | $1.1248\times10^7$ | 15.7555 |
| | Power | $\hat{\beta}$ | -0.5135 | 70515 | $1.0138\times10^8$ | 16.2532 |
| X | Exponential | $\hat{\alpha}$ | 0.0159 | 234026 | $2.9523\times10^7$ | 16.309 |
| | Power | $\hat{\beta}$ | -0.5238 | 650033 | $6.7308\times10^8$ | 24.9823 |

Table A.3: Score function data for Chromosomes 16 to 22 and X's estimated configurations, found using metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $\chi^2$ (4.9) value it minimizes. Column six and seven give the SSR$(\mathbf{M}, \hat{\mathbf{U}})$ (4.8) and $S_3(\hat{\mathbf{X}})$ (2.14) values found using the estimated parameter values. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome's Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 2$ adjustment, and fitting into three dimensional space with metric MDS. The statistics $\chi^2$, SSR$(\mathbf{M}, \hat{\mathbf{U}})$ and $S_3(\hat{\mathbf{X}})$ are then extracted from the fitted configuration.

## A.2 Chromosome score function data from non-metric MDS

Appendix containing the score function data when obtaining the estimated chromosome configurations using the exponential transform (4.4) or the power transform (4.6), and using non-metric MDS.

| Chromosome | Transform function | Parameter | Parameter estimate | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|
| 1 | Exponential | $\hat{\alpha}$ | 0.1211 | 13.6395 |
| | Power | $\hat{\beta}$ | -0.1686 | 13.6552 |
| 2 | Exponential | $\hat{\alpha}$ | 0.1775 | 14.4785 |
| | Power | $\hat{\beta}$ | -0.1701 | 14.4759 |
| 3 | Exponential | $\hat{\alpha}$ | 0.1841 | 13.3954 |
| | Power | $\hat{\beta}$ | -0.128 | 13.3988 |
| 4 | Exponential | $\hat{\alpha}$ | 0.1111 | 13.0859 |
| | Power | $\hat{\beta}$ | -0.1948 | 13.0849 |
| 5 | Exponential | $\hat{\alpha}$ | 0.1019 | 12.3782 |
| | Power | $\hat{\beta}$ | -0.2725 | 12.3432 |
| 6 | Exponential | $\hat{\alpha}$ | 0.1238 | 13.4278 |
| | Power | $\hat{\beta}$ | -0.343 | 13.4295 |
| 7 | Exponential | $\hat{\alpha}$ | 0.2502 | 15.9983 |
| | Power | $\hat{\beta}$ | -0.1851 | 14.702 |

Table A.4: Score function data for Chromosomes 1 to 7's estimated configurations, found using non-metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $S_3(\hat{\mathbf{X}})$ (2.14) value it minimizes. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 1$ adjustment, and fitting into three dimensional space with non-metric MDS. The statistic $S_3(\hat{\mathbf{X}})$ is then extracted from the fitted configuration.

| Chromosome | Transform function | Parameter | Parameter estimate | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|
| 8 | Exponential | $\hat{\alpha}$ | 0.4294 | 12.5188 |
| | Power | $\hat{\beta}$ | -0.3853 | 12.4885 |
| 9 | Exponential | $\hat{\alpha}$ | 0.2211 | 12.3408 |
| | Power | $\hat{\beta}$ | -0.154 | 12.3598 |
| 10 | Exponential | $\hat{\alpha}$ | 0.1877 | 13.2302 |
| | Power | $\hat{\beta}$ | -0.3083 | 13.2288 |
| 11 | Exponential | $\hat{\alpha}$ | 0.1533 | 12.676 |
| | Power | $\hat{\beta}$ | -0.1486 | 12.6801 |
| 12 | Exponential | $\hat{\alpha}$ | 0.3034 | 13.9874 |
| | Power | $\hat{\beta}$ | -0.2616 | 12.9924 |
| 13 | Exponential | $\hat{\alpha}$ | 0.0739 | 13.5248 |
| | Power | $\hat{\beta}$ | -0.3085 | 13.5243 |
| 14 | Exponential | $\hat{\alpha}$ | 0.0961 | 12.1728 |
| | Power | $\hat{\beta}$ | -0.304 | 12.1724 |
| 15 | Exponential | $\hat{\alpha}$ | 0.0412 | 12.2888 |
| | Power | $\hat{\beta}$ | -0.3157 | 12.2927 |

Table A.5: Score function data for Chromosomes 1 to 7's estimated configurations, found using non-metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $S_3(\hat{\mathbf{X}})$ (2.14) value it minimizes. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 1$ adjustment, and fitting into three dimensional space with non-metric MDS. The statistic $S_3(\hat{\mathbf{X}})$ is then extracted from the fitted configuration.

| Chromosome | Transform function | Parameter | Parameter estimate | $S_3(\hat{\mathbf{X}})\%$ |
|---|---|---|---|---|
| 16 | Exponential | $\hat{\alpha}$ | 0.0745 | 12.5066 |
| | Power | $\hat{\beta}$ | -0.309 | 12.5109 |
| 17 | Exponential | $\hat{\alpha}$ | 0.073 | 12.9384 |
| | Power | $\hat{\beta}$ | -0.1102 | 12.9348 |
| 18 | Exponential | $\hat{\alpha}$ | 0.3802 | 19.0418 |
| | Power | $\hat{\beta}$ | -0.2307 | 13.82 |
| 19 | Exponential | $\hat{\alpha}$ | 0.0319 | 10.8232 |
| | Power | $\hat{\beta}$ | -0.2366 | 10.8267 |
| 20 | Exponential | $\hat{\alpha}$ | 0.2165 | 12.6557 |
| | Power | $\hat{\beta}$ | -0.4098 | 10.9737 |
| 21 | Exponential | $\hat{\alpha}$ | 0.1515 | 13.6104 |
| | Power | $\hat{\beta}$ | -0.3052 | 9.8415 |
| 22 | Exponential | $\hat{\alpha}$ | 0.3033 | 15.4072 |
| | Power | $\hat{\beta}$ | -0.2834 | 9.1781 |
| X | Exponential | $\hat{\alpha}$ | 0.1811 | 12.4171 |
| | Power | $\hat{\beta}$ | -0.2093 | 12.4245 |

Table A.6: Score function data for Chromosomes 1 to 7's estimated configurations, found using non-metric MDS. Column one lists the chromosome the row is referring to. Column two and three list which transform function has been used and which parameter has been estimated. Column four and five give the estimated parameter value and the $S_3(\hat{\mathbf{X}})$ (2.14) value it minimizes. The data is found by applying the fitting algorithm (Section 4.1.3) to the chromosome Hi-C count matrix, with either the exponential transform (4.4) or power transform (4.6) with the $m_{\text{min}} = 1$ adjustment, and fitting into three dimensional space with non-metric MDS. The statistic $S_3(\hat{\mathbf{X}})$ is then extracted from the fitted configuration.

# B Model based approach simulation results

Appendix containing model-based approach (MBA) simulation results, to compliment results in Section 5.2.

## B.1 Metric multidimensional scaling results

Simulation results using the MBA found with metric multidimensional scaling (MDS).

| $\theta_{1:p}$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 17.5280 | 13.7420 | 11.0115 | 9.0516 | 19.2844 | 15.6954 | 13.1714 | 11.4195 |
| 0.01 | 40.7642 | 32.8719 | 25.9647 | 20.2632 | 42.0806 | 34.2533 | 27.4567 | 21.9221 |
| 0.001 | 68.5834 | 60.7124 | 52.2316 | 43.6602 | 69.5229 | 61.7720 | 53.4092 | 44.9184 |
| 0.0001 | 87.3587 | 82.9926 | 77.5445 | 70.9412 | 87.8165 | 83.6021 | 78.2862 | 71.8540 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.0085 | 19.6464 | 16.4867 | 14.1867 | 32.2633 | 26.4798 | 21.8590 | 18.1601 |
| 0.01 | 49.5337 | 41.2414 | 33.6520 | 27.1904 | 60.3159 | 52.0161 | 43.7954 | 36.1912 |
| 0.001 | 75.6007 | 68.6874 | 60.8670 | 52.4629 | 82.7933 | 77.3000 | 70.6903 | 63.1057 |
| 0.0001 | 90.7337 | 87.3872 | 83.0481 | 77.6207 | 93.8314 | 91.4926 | 88.3829 | 84.3275 |

Table B.7: Percentage of information projected into the first $k$ dimensions $\theta_{1:p}$ (2.11) from the MBA simulations, found using the exponential transform (4.4) with metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$.

| $\theta_{1:p}$ | **Power transform** $m_{\text{min}} = 1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 27.8404 | 20.1293 | 13.6644 | 9.8202 | 30.7130 | 24.7237 | 19.1479 | 14.3174 |
| 0.01 | 57.6572 | 48.9282 | 40.1297 | 31.7891 | 58.7113 | 50.1035 | 41.5921 | 33.8346 |
| 0.001 | 81.2577 | 75.3700 | 68.3916 | 60.4241 | 81.8946 | 76.1679 | 69.3219 | 61.5269 |
| 0.0001 | 93.2063 | 90.6505 | 87.2819 | 82.9114 | 93.4746 | 91.0144 | 87.7434 | 83.4877 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.9252 | 18.8831 | 13.7386 | 10.9278 | 21.9326 | 14.8751 | 10.4338 | 9.13840 |
| 0.01 | 54.6591 | 45.8674 | 37.1914 | 29.0098 | 52.1325 | 43.3325 | 34.7716 | 26.7928 |
| 0.001 | 79.3910 | 73.1468 | 65.7938 | 57.5829 | 77.7440 | 71.1615 | 63.5305 | 55.1377 |
| 0.0001 | 92.4360 | 89.6187 | 85.9195 | 81.1791 | 91.7041 | 88.6540 | 84.6881 | 79.6137 |
| $\theta_{1:p}$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 27.8637 | 20.4204 | 14.6388 | 11.1370 | 30.7036 | 24.2488 | 18.6172 | 14.6818 |
| 0.01 | 57.5900 | 48.9094 | 40.1247 | 31.8003 | 58.6961 | 50.1368 | 41.5896 | 33.8635 |
| 0.001 | 81.2557 | 75.4122 | 68.4265 | 60.4729 | 81.9058 | 76.1921 | 69.3358 | 61.4963 |
| 0.0001 | 93.2115 | 90.6612 | 87.2650 | 82.9057 | 93.4705 | 91.0086 | 87.7531 | 83.4964 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.8325 | 18.5495 | 14.3731 | 12.0332 | 22.1934 | 16.0361 | 12.6126 | 11.4270 |
| 0.01 | 54.6531 | 45.8829 | 37.2032 | 29.0013 | 52.2750 | 43.3092 | 34.7572 | 26.8080 |
| 0.001 | 79.3953 | 73.1522 | 65.7945 | 57.5684 | 77.8475 | 71.1560 | 63.5265 | 55.1326 |
| 0.0001 | 92.4297 | 89.6134 | 85.9065 | 81.1836 | 91.6675 | 88.6615 | 84.6817 | 79.6197 |

Table B.8: Percentage of information projected into the first $k$ dimensions $\theta_{1:p}$ (2.11) from the MBA simulations, found using the power transform (4.6) with $\beta = -0.5$ and with metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0690 | 0.0930 | 0.1224 | 0.1579 | 0.2271 | 0.2666 | 0.3045 | 0.3495 |
| 0.01 | 0.0212 | 0.0229 | 0.0418 | 0.0577 | 0.0800 | 0.1112 | 0.1506 | 0.1962 |
| 0.001 | 0.0067 | 0.0095 | 0.0134 | 0.0189 | 0.0252 | 0.3555 | 0.5020 | 0.0708 |
| 0.0001 | 0.0021 | 0.0030 | 0.0042 | 0.0060 | 0.0080 | 0.0112 | 0.0160 | 0.0225 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.1901 | 0.1430 | 0.1831 | 0.2276 | 0.0510 | 0.0674 | 0.0860 | 0.1084 |
| 0.01 | 0.0343 | 0.0480 | 0.0668 | 0.0913 | 0.0160 | 0.0224 | 0.0312 | 0.0427 |
| 0.001 | 0.0108 | 0.0153 | 0.0216 | 0.0304 | 0.0051 | 0.0072 | 0.0101 | 0.0142 |
| 0.0001 | 0.0034 | 0.0049 | 0.0069 | 0.0097 | 0.0016 | 0.0023 | 0.0032 | 0.0045 |

Table B.9: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the MBA simulations, found using the exponential transform (4.4) with metric MDS.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0665 | 0.1319 | 0.2559 | 0.3420 | 0.3996 | 0.5022 | 0.6112 | 0.6561 |
| 0.01 | 0.0158 | 0.0228 | 0.0331 | 0.0497 | 0.0649 | 0.1002 | 0.1944 | 0.3375 |
| 0.001 | 0.0049 | 0.0070 | 0.0099 | 0.0141 | 0.0193 | 0.0273 | 0.0392 | 0.0570 |
| 0.0001 | 0.0015 | 0.0022 | 0.0031 | 0.0044 | 0.0060 | 0.0085 | 0.0121 | 0.0171 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2422 | 0.5582 | 0.6301 | 0.6415 | 0.0981 | 0.1742 | 0.2564 | 0.2749 |
| 0.01 | 0.0380 | 0.0544 | 0.0808 | 0.1267 | 0.0223 | 0.0319 | 0.0461 | 0.0681 |
| 0.001 | 0.0117 | 0.0166 | 0.0235 | 0.0334 | 0.0070 | 0.0098 | 0.0139 | 0.0198 |
| 0.0001 | 0.0037 | 0.0052 | 0.0074 | 0.0104 | 0.0022 | 0.0031 | 0.0044 | 0.0062 |
| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0651 | 0.1093 | 0.1767 | 0.2241 | 0.3991 | 0.4797 | 0.5436 | 0.5739 |
| 0.01 | 0.0160 | 0.0227 | 0.0333 | 0.0496 | 0.0648 | 0.1004 | 0.1980 | 0.3367 |
| 0.001 | 0.0049 | 0.0070 | 0.0098 | 0.0140 | 0.0192 | 0.0275 | 0.0392 | 0.0565 |
| 0.0001 | 0.0016 | 0.0022 | 0.0031 | 0.0044 | 0.0060 | 0.0086 | 0.0120 | 0.0171 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2151 | 0.4604 | 0.5394 | 0.5471 | 0.0913 | 0.1388 | 0.1821 | 0.1999 |
| 0.01 | 0.0379 | 0.0544 | 0.0810 | 0.1277 | 0.0228 | 0.0319 | 0.0461 | 0.0681 |
| 0.001 | 0.0117 | 0.0165 | 0.0235 | 0.0335 | 0.0069 | 0.0098 | 0.0139 | 0.0198 |
| 0.0001 | 0.0037 | 0.0052 | 0.0074 | 0.0105 | 0.0022 | 0.0031 | 0.0044 | 0.0062 |

Table B.10: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the MBA simulations, found using the power transform (4.6) with metric MDS. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

| $G(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0247 | 1.0436 | 1.0702 | 1.0977 | 1.0595 | 1.0977 | 1.1492 | 1.2068 |
| 0.01 | 1.0026 | 1.0051 | 1.0101 | 1.0190 | 1.0071 | 1.0136 | 1.0256 | 1.0457 |
| 0.001 | 1.0002 | 1.0005 | 1.0010 | 1.0021 | 1.0007 | 1.0014 | 1.0029 | 1.0057 |
| 0.0001 | 1.0000 | 1.0001 | 1.0001 | 1.0002 | 1.0001 | 1.0001 | 1.0003 | 1.0006 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0228 | 1.0384 | 1.0589 | 1.0765 | 0.9949 | 0.9890 | 0.9765 | 0.9513 |
| 0.01 | 1.0025 | 1.0047 | 1.0093 | 1.0173 | 0.9995 | 0.9991 | 0.9981 | 0.9961 |
| 0.001 | 1.0002 | 1.0005 | 1.0010 | 1.0020 | 1.0000 | 0.9999 | 0.9998 | 0.9996 |
| 0.0001 | 1.0000 | 1.0001 | 1.0001 | 1.0002 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Table B.11: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the MBA simulations, found using the exponential transform (4.4) with metric MDS.

| $G(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0375 | 1.0854 | 1.2083 | 1.4712 | 1.0758 | 1.1875 | 1.4101 | 1.7622 |
| 0.01 | 1.0031 | 1.0065 | 1.0131 | 1.0268 | 1.0041 | 1.0085 | 1.0190 | 1.0461 |
| 0.001 | 1.0003 | 1.0006 | 1.0013 | 1.0027 | 1.0004 | 1.0008 | 1.0016 | 1.0033 |
| 0.0001 | 1.0000 | 1.0001 | 1.0001 | 1.0003 | 1.0000 | 1.0001 | 1.0002 | 1.0003 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0583 | 1.1776 | 1.3932 | 1.7370 | 1.0668 | 1.1533 | 1.3524 | 1.6763 |
| 0.01 | 1.0041 | 1.0081 | 1.0166 | 1.0351 | 1.0054 | 1.0108 | 1.0217 | 1.0452 |
| 0.001 | 1.0004 | 1.0008 | 1.0016 | 1.0032 | 1.0005 | 1.0011 | 1.0021 | 1.0042 |
| 0.0001 | 1.0000 | 1.0001 | 1.0002 | 1.0003 | 1.0001 | 1.0001 | 1.0002 | 1.0004 |

| $G(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0370 | 1.0805 | 1.1766 | 1.3559 | 1.0750 | 1.1635 | 1.3156 | 1.5425 |
| 0.01 | 1.0033 | 1.0065 | 1.0132 | 1.0271 | 1.0041 | 1.0085 | 1.0190 | 1.0463 |
| 0.001 | 1.0003 | 1.0006 | 1.0013 | 1.0026 | 1.0004 | 1.0008 | 1.0016 | 1.0032 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0002 | 1.0000 | 1.0001 | 1.0002 | 1.0003 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0524 | 1.1316 | 1.2815 | 1.4901 | 1.0634 | 1.1404 | 1.2755 | 1.4654 |
| 0.01 | 1.0041 | 1.0081 | 1.0167 | 1.0349 | 1.0047 | 1.0108 | 1.0220 | 1.0451 |
| 0.001 | 1.0004 | 1.0008 | 1.0016 | 1.0032 | 1.0006 | 1.0010 | 1.0021 | 1.0043 |
| 0.0001 | 1.0000 | 1.0001 | 1.0002 | 1.0003 | 1.0001 | 1.0001 | 1.0002 | 1.0004 |

Table B.12: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the MBA simulations, found using the power transform (4.6) with metric MDS. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

## B.2 Non-metric multidimensional scaling results

Simulation results using the MBA found with non-metric multidimensional scaling (MDS).

| $S_p(\hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 21.7476 | 27.9854 | 34.2747 | 39.7239 | 17.3766 | 22.4655 | 27.5527 | 31.5850 |
| 0.01 | 7.2434 | 10.1676 | 14.0784 | 19.0992 | 6.0564 | 8.4388 | 11.6124 | 15.6956 |
| 0.001 | 2.2221 | 3.1820 | 4.5327 | 6.4265 | 2.1430 | 2.8246 | 3.8657 | 5.4252 |
| 0.0001 | 0.6055 | 0.9362 | 1.3713 | 1.9764 | 0.6426 | 0.9296 | 1.3404 | 1.9222 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 15.5261 | 20.3071 | 25.1777 | 29.2005 | 12.8483 | 16.8108 | 20.8579 | 24.3327 |
| 0.01 | 5.2815 | 7.3932 | 10.2086 | 13.9151 | 4.1562 | 5.8554 | 8.1614 | 11.1793 |
| 0.001 | 1.6728 | 2.3873 | 3.3932 | 4.7216 | 1.2475 | 1.8133 | 2.5965 | 3.7034 |
| 0.0001 | 0.4234 | 0.7044 | 1.0553 | 1.5343 | 0.1021 | 0.3540 | 0.6976 | 1.0985 |

Table B.13: Stress of fit values $S_p(\hat{\mathbf{X}})$ (2.14) from the MBA simulations, found using the exponential transform (4.4) with non-metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$.

| $S_p(\hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 8.1963 | 10.6937 | 13.6593 | 17.0108 | 4.0481 | 5.3433 | 6.9636 | 8.9428 |
| 0.01 | 2.8595 | 3.9825 | 5.4632 | 7.3621 | 1.4809 | 2.0252 | 2.7437 | 3.6832 |
| 0.001 | 0.8581 | 1.2528 | 1.7989 | 2.5489 | 0.6983 | 0.7118 | 0.9669 | 1.3361 |
| 0.0001 | 0.1667 | 0.3099 | 0.4989 | 0.7560 | 0.2248 | 0.3271 | 0.4741 | 0.6528 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 5.4050 | 6.9951 | 8.9464 | 11.2338 | 8.5694 | 10.9940 | 13.7361 | 16.6639 |
| 0.01 | 2.1350 | 2.8526 | 3.7786 | 4.9568 | 3.3417 | 4.5238 | 6.0336 | 7.9233 |
| 0.001 | 0.7496 | 1.0422 | 1.4297 | 1.9364 | 1.1241 | 1.5842 | 2.2018 | 3.0278 |
| 0.0001 | 0.2096 | 0.3194 | 0.4703 | 0.6719 | 0.3113 | 0.4725 | 0.6972 | 1.0050 |

| $S_p(\hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 8.1891 | 10.6967 | 13.6682 | 17.0050 | 4.0488 | 5.3459 | 6.9677 | 8.9452 |
| 0.01 | 2.8609 | 3.9811 | 5.4593 | 7.3565 | 1.4810 | 2.0258 | 2.7439 | 3.6817 |
| 0.001 | 0.8573 | 1.2510 | 1.7998 | 2.5487 | 0.6962 | 0.7145 | 0.9687 | 1.3361 |
| 0.0001 | 0.1669 | 0.3101 | 0.4989 | 0.7567 | 0.2232 | 0.3271 | 0.4814 | 0.6480 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 5.4056 | 6.9980 | 8.9485 | 11.2341 | 8.5684 | 11.0089 | 13.7463 | 16.6624 |
| 0.01 | 2.1339 | 2.8544 | 3.7772 | 4.9594 | 3.3417 | 4.5238 | 6.0336 | 7.9233 |
| 0.001 | 0.7491 | 1.0435 | 1.4305 | 1.9383 | 1.1241 | 1.5842 | 2.2018 | 3.0278 |
| 0.0001 | 0.2100 | 0.3194 | 0.4700 | 0.6717 | 0.3113 | 0.4725 | 0.6972 | 1.0050 |

Table B.14: Stress of fit values $S_p(\hat{\mathbf{X}})$ (2.14) from the MBA simulations, found using the power transform (4.6) with non-metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0356 | 0.0425 | 0.0551 | 0.0752 | 0.4663 | 0.4961 | 0.5247 | 0.5270 |
| 0.01 | 0.0111 | 0.0153 | 0.0212 | 0.0286 | 0.3527 | 0.3798 | 0.4126 | 0.4459 |
| 0.001 | 0.0034 | 0.0049 | 0.0069 | 0.0096 | 0.0549 | 0.2376 | 0.3147 | 0.3423 |
| 0.0001 | 0.0010 | 0.0015 | 0.0022 | 0.0031 | 0.0275 | 0.0290 | 0.0339 | 0.0432 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2672 | 0.2969 | 0.3319 | 0.3607 | 0.0644 | 0.0849 | 0.1283 | 0.2149 |
| 0.01 | 0.1610 | 0.1832 | 0.2122 | 0.2408 | 0.0134 | 0.0183 | 0.0253 | 0.0355 |
| 0.001 | 0.0819 | 0.0901 | 0.0988 | 0.1474 | 0.0039 | 0.0056 | 0.0079 | 0.0118 |
| 0.0001 | 0.0034 | 0.0046 | 0.0058 | 0.0094 | 0.0003 | 0.0011 | 0.0022 | 0.0034 |

Table B.15: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the MBA simulations, found using the exponential transform (4.4) with non-metric MDS.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 1$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0204 | 0.0276 | 0.0358 | 0.0435 | 0.3909 | 0.4172 | 0.4450 | 0.4732 |
| 0.01 | 0.0051 | 0.0077 | 0.0114 | 0.0167 | 0.3153 | 0.3350 | 0.3570 | 0.3814 |
| 0.001 | 0.0014 | 0.0020 | 0.0029 | 0.0044 | 0.1255 | 0.2682 | 0.2920 | 0.3092 |
| 0.0001 | 0.0003 | 0.0005 | 0.0008 | 0.0012 | 0.0492 | 0.0618 | 0.0694 | 0.0997 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2445 | 0.2732 | 0.3029 | 0.3333 | 0.0230 | 0.0286 | 0.0343 | 0.0408 |
| 0.01 | 0.1586 | 0.1819 | 0.2073 | 0.2344 | 0.0095 | 0.0123 | 0.0158 | 0.0203 |
| 0.001 | 0.0977 | 0.1134 | 0.1312 | 0.1514 | 0.0032 | 0.0045 | 0.0062 | 0.0084 |
| 0.0001 | 0.0546 | 0.0674 | 0.0796 | 0.0929 | 0.0009 | 0.0014 | 0.0021 | 0.0029 |
| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0203 | 0.0276 | 0.0359 | 0.0436 | 0.3908 | 0.4174 | 0.4449 | 0.4732 |
| 0.01 | 0.0051 | 0.0076 | 0.0114 | 0.0166 | 0.3152 | 0.3350 | 0.3570 | 0.3815 |
| 0.001 | 0.0013 | 0.0020 | 0.0030 | 0.0044 | 0.1262 | 0.2674 | 0.2917 | 0.3093 |
| 0.0001 | 0.0003 | 0.0005 | 0.0008 | 0.0012 | 0.0523 | 0.0609 | 0.0613 | 0.1044 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2444 | 0.2734 | 0.3032 | 0.3336 | 0.0299 | 0.0288 | 0.0350 | 0.0403 |
| 0.01 | 0.1587 | 0.1818 | 0.2072 | 0.2346 | 0.0095 | 0.0123 | 0.0158 | 0.0203 |
| 0.001 | 0.0977 | 0.1135 | 0.1312 | 0.1514 | 0.0032 | 0.0045 | 0.0062 | 0.0084 |
| 0.0001 | 0.0544 | 0.0673 | 0.0797 | 0.0930 | 0.0009 | 0.0014 | 0.0021 | 0.0029 |

Table B.16: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the MBA simulations, found using the power transform (4.6) with non-metric MDS. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

# C Unbiased MBA simulation results

Appendix containing unbiased model-based approach (MBA) simulation results, to compliment results in Section 5.4.

## C.1 Metric multidimensional scaling

Simulation results using the unbiased MBA found with metric multidimensional scaling (MDS).

| $\theta_{1:p}\%$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 18.1232 | 13.7255 | 10.4529 | 8.1772 | 19.7401 | 15.5430 | 12.4650 | 10.3871 |
| 0.01 | 40.8812 | 32.8552 | 25.7538 | 19.7849 | 42.1234 | 34.1592 | 27.1522 | 21.3354 |
| 0.001 | 68.5887 | 60.6952 | 52.2197 | 43.5959 | 69.5326 | 61.7606 | 53.3779 | 44.8199 |
| 0.0001 | 87.3442 | 82.9960 | 77.5358 | 70.9382 | 87.8156 | 83.5995 | 78.2852 | 71.8358 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.0104 | 18.5586 | 14.3490 | 11.3701 | 32.2346 | 24.9538 | 18.6919 | 13.5314 |
| 0.01 | 49.5419 | 41.0156 | 33.0446 | 26.0161 | 60.3183 | 51.7872 | 43.1059 | 34.7738 |
| 0.001 | 75.5989 | 68.6590 | 60.7976 | 52.3137 | 82.7878 | 77.2864 | 70.6369 | 62.9474 |
| 0.0001 | 90.7331 | 87.3790 | 83.0363 | 77.5971 | 93.8327 | 91.4944 | 88.3834 | 84.3287 |

Table C.17: Percentage of information projected into the first $p$ dimensions $\theta_{1:p}$ (2.11) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the exponential transform (5.21) with metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$.

| $\theta_{1:p}\%$ | **Power transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 30.5899 | 24.0084 | 18.6904 | 14.6900 | 32.5093 | 26.4379 | 21.5617 | 17.9681 |
| 0.01 | 57.8888 | 49.3147 | 40.8278 | 32.8846 | 58.9830 | 50.5663 | 42.2604 | 34.7111 |
| 0.001 | 81.2722 | 75.4005 | 68.4488 | 60.5409 | 81.9266 | 76.2179 | 69.4123 | 61.6248 |
| 0.0001 | 93.2081 | 90.6522 | 87.2742 | 82.9115 | 83.4725 | 91.0162 | 87.7375 | 83.4963 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 28.3918 | 22.3629 | 17.6879 | 14.5716 | 26.3759 | 20.6088 | 16.1115 | 13.0099 |
| 0.01 | 55.0399 | 46.4456 | 38.1567 | 30.6154 | 52.5690 | 44.0336 | 35.8399 | 28.4991 |
| 0.001 | 79.4252 | 73.1882 | 65.9106 | 57.7595 | 77.7691 | 71.2159 | 63.6485 | 55.3393 |
| 0.0001 | 92.4243 | 89.6178 | 85.9176 | 81.1940 | 91.6995 | 88.6632 | 84.6755 | 79.6224 |

Table C.18: Percentage of information projected into the first $p$ dimensions $\theta_{1:p}$ (2.11) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the power transform (5.22) with metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 0.0667 | 0.0947 | 0.1334 | 0.1858 | 0.2630 | 0.3382 | 0.3891 | 0.4258 |
| 0.01 | 0.0212 | 0.0299 | 0.0424 | 0.0597 | 0.0805 | 0.1146 | 0.1636 | 0.2353 |
| 0.001 | 0.0067 | 0.0095 | 0.0134 | 0.0189 | 0.0252 | 0.0357 | 0.0507 | 0.0719 |
| 0.0001 | 0.0021 | 0.0003 | 0.0042 | 0.0006 | 0.0080 | 0.0113 | 0.0159 | 0.0225 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 0.1087 | 0.1526 | 0.2103 | 0.2839 | 0.0513 | 0.0736 | 0.1072 | 0.1594 |
| 0.01 | 0.0343 | 0.0484 | 0.0686 | 0.0970 | 0.0160 | 0.0227 | 0.0321 | 0.0456 |
| 0.001 | 0.0108 | 0.0153 | 0.0217 | 0.0306 | 0.0051 | 0.0072 | 0.0101 | 0.0143 |
| 0.0001 | 0.0034 | 0.0048 | 0.0068 | 0.0097 | 0.0016 | 0.0023 | 0.0032 | 0.0045 |

Table C.19: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the exponential transform (5.21) with metric MDS.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0492 | 0.0694 | 0.0979 | 0.1366 | 0.0302 | 0.3824 | 0.4453 | 0.5035 |
| 0.01 | 0.0156 | 0.0220 | 0.0310 | 0.0438 | 0.0629 | 0.0939 | 0.1514 | 0.2634 |
| 0.001 | 0.0049 | 0.0070 | 0.0098 | 0.0140 | 0.0192 | 0.0271 | 0.0388 | 0.0557 |
| 0.0001 | 0.0015 | 0.0022 | 0.0031 | 0.0044 | 0.0060 | 0.0085 | 0.0122 | 0.0171 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.1208 | 0.1790 | 0.2888 | 0.4739 | 0.0687 | 0.0950 | 0.1305 | 0.1719 |
| 0.01 | 0.3690 | 0.0524 | 0.0747 | 0.1068 | 0.0219 | 0.0310 | 0.0436 | 0.0613 |
| 0.001 | 0.0116 | 0.0166 | 0.0232 | 0.0331 | 0.0070 | 0.0098 | 0.0139 | 0.0197 |
| 0.0001 | 0.0037 | 0.0052 | 0.0073 | 0.0104 | 0.0022 | 0.0031 | 0.0044 | 0.0062 |

Table C.20: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the power transform (5.22) with metric MDS.

| $G(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0139 | 1.0274 | 1.0571 | 1.1205 | 1.0441 | 1.0789 | 1.1391 | 1.2456 |
| 0.01 | 1.0013 | 1.0028 | 1.0056 | 1.0114 | 1.0050 | 1.0100 | 1.0193 | 1.0364 |
| 0.001 | 1.0001 | 1.0003 | 1.0006 | 1.0011 | 1.0005 | 1.0010 | 1.0021 | 1.0041 |
| 0.0001 | 1.0000 | 1.0000 | 1.0000 | 1.0001 | 1.0001 | 1.0001 | 1.0002 | 1.0004 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0146 | 1.0286 | 1.0567 | 1.1140 | 0.9983 | 0.9963 | 0.9916 | 0.9849 |
| 0.01 | 1.0015 | 1.0030 | 1.0061 | 1.0120 | 0.9999 | 0.9997 | 0.9995 | 0.9987 |
| 0.001 | 1.0002 | 1.0003 | 1.0006 | 1.0013 | 1.0000 | 1.0000 | 0.9999 | 0.9999 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |

Table C.21: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the exponential transform (5.21) with metric MDS.

| $G(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0091 | 1.0177 | 1.0355 | 1.0701 | 1.0279 | 1.0548 | 1.0987 | 1.1702 |
| 0.01 | 1.0009 | 1.0017 | 1.0035 | 1.0071 | 1.0024 | 1.0049 | 1.0102 | 1.0218 |
| 0.001 | 1.0001 | 1.0002 | 1.0003 | 1.0006 | 1.0002 | 1.0005 | 1.0010 | 1.0019 |
| 0.0001 | 1.0000 | 1.0000 | 1.0000 | 1.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0002 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0148 | 1.0303 | 1.0612 | 1.1284 | 1.0146 | 1.0291 | 1.0578 | 1.1161 |
| 0.01 | 1.0015 | 1.0029 | 1.0059 | 1.0119 | 1.0015 | 1.0029 | 1.0059 | 1.0119 |
| 0.001 | 1.0001 | 1.0003 | 1.0006 | 1.0012 | 1.0001 | 1.0003 | 1.0006 | 1.0012 |
| 0.0001 | 1.0000 | 1.0000 | 1.0000 | 1.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 |

Table C.22: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the power transform (5.22) with metric MDS.

## C.2 Non-metric multidimensional scaling results

Simulation results using the unbiased MBA found with non-metric multidimensional scaling (MDS).

| $S_p(\hat{\mathbf{X}})\%$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 21.2432 | 27.5824 | 33.5729 | 38.2428 | 17.6751 | 22.7543 | 26.9747 | 29.6265 |
| 0.01 | 7.1929 | 10.1345 | 14.1317 | 19.3431 | 6.0609 | 8.4646 | 11.7544 | 16.0850 |
| 0.001 | 2.2227 | 3.1806 | 4.5342 | 6.4307 | 2.1370 | 2.8197 | 3.8641 | 5.4313 |
| 0.0001 | 0.6046 | 0.9366 | 1.3720 | 1.9783 | 0.6425 | 0.9287 | 1.3413 | 1.9170 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 15.4024 | 20.1307 | 24.7785 | 28.3750 | 12.4512 | 16.5272 | 20.9157 | 25.1266 |
| 0.01 | 5.2625 | 7.3685 | 10.2080 | 13.9861 | 4.1522 | 5.8454 | 8.1826 | 11.2910 |
| 0.001 | 1.6694 | 2.3822 | 3.3850 | 4.7205 | 1.2475 | 1.8120 | 2.5972 | 3.7110 |
| 0.0001 | 0.4244 | 0.7038 | 1.0555 | 1.5339 | 0.1015 | 0.3543 | 0.6989 | 1.0983 |

Table C.23: Stress of fit values $S_p(\hat{\mathbf{X}})$ (2.14) from the MBA simulations, found using unbiased perturbed distances emulating the exponential transform (5.21) with non-metric MDS. For the straight line $p=1$ and for the parabola; semi-circle and circle $p=2$.

| $S_p(\hat{\mathbf{X}})\%$ | **Power transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 8.3830 | 11.3437 | 15.3389 | 20.5634 | 4.2018 | 5.7649 | 8.0983 | 11.7332 |
| 0.01 | 2.8609 | 4.0064 | 5.5407 | 7.5909 | 1.4911 | 2.0461 | 2.7934 | 3.8078 |
| 0.001 | 0.8574 | 1.2517 | 1.8000 | 2.5593 | 0.7381 | 0.7457 | 0.9717 | 1.3428 |
| 0.0001 | 0.1670 | 0.3102 | 0.4988 | 0.7568 | 0.2365 | 0.3600 | 0.5271 | 0.7475 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 5.7965 | 7.9396 | 11.2415 | 16.0666 | 9.5325 | 12.9781 | 17.7032 | 23.0729 |
| 0.01 | 2.1596 | 2.9033 | 3.9205 | 5.2447 | 3.4118 | 4.6796 | 6.3770 | 8.6593 |
| 0.001 | 0.7553 | 1.0703 | 1.4369 | 1.9666 | 1.1286 | 1.5936 | 2.2262 | 3.0791 |
| 0.0001 | 0.2116 | 0.3202 | 0.4741 | 0.7078 | 0.3117 | 0.4732 | 0.6986 | 1.0070 |

Table C.24: Stress of fit values $S_p(\hat{\mathbf{X}})$ (2.14) from the MBA simulations, found using unbiased perturbed distances emulating the power transform (5.22) with non-metric MDS. For the straight line $p=1$ and for the parabola; semi-circle and circle $p=2$.

| $P(\mathbf{X},\hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 0.0331 | 0.0428 | 0.0548 | 0.0668 | 0.4406 | 0.4717 | 0.5127 | 0.5538 |
| 0.01 | 0.0108 | 0.0150 | 0.0213 | 0.0300 | 0.3500 | 0.3771 | 0.4070 | 0.4334 |
| 0.001 | 0.0034 | 0.0048 | 0.0069 | 0.0096 | 0.0586 | 0.2390 | 0.3154 | 0.3428 |
| 0.0001 | 0.0010 | 0.0015 | 0.0022 | 0.0031 | 0.0278 | 0.0298 | 0.0330 | 0.0480 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ | $\rho=1$ | $\rho=2$ | $\rho=4$ | $\rho=8$ |
| 0.1 | 0.2438 | 0.2747 | 0.3244 | 0.3796 | 0.0396 | 0.0561 | 0.0789 | 0.1068 |
| 0.01 | 0.1575 | 0.1818 | 0.2089 | 0.2353 | 0.0132 | 0.0175 | 0.0247 | 0.0355 |
| 0.001 | 0.0817 | 0.0935 | 0.1039 | 0.1487 | 0.0039 | 0.0055 | 0.0079 | 0.0119 |
| 0.0001 | 0.0032 | 0.0046 | 0.0063 | 0.0098 | 0.0003 | 0.0011 | 0.0022 | 0.0034 |

Table C.25: Shape difference values $P(\mathbf{X},\hat{\mathbf{X}})$ (5.5) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the exponential transform (5.21) with non-metric MDS.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0192 | 0.0273 | 0.0361 | 0.0439 | 0.3945 | 0.4269 | 0.4665 | 0.5124 |
| 0.01 | 0.0051 | 0.0076 | 0.0114 | 0.0171 | 0.3150 | 0.3354 | 0.3587 | 0.3850 |
| 0.001 | 0.0013 | 0.0020 | 0.0029 | 0.0044 | 0.1162 | 0.2512 | 0.2917 | 0.3094 |
| 0.0001 | 0.0003 | 0.0005 | 0.0008 | 0.0012 | 0.0185 | 0.0056 | 0.0078 | 0.0273 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.2524 | 0.2902 | 0.3358 | 0.3831 | 0.0242 | 0.0331 | 0.0485 | 0.0749 |
| 0.01 | 0.1580 | 0.1825 | 0.2092 | 0.2408 | 0.0097 | 0.0127 | 0.0166 | 0.0219 |
| 0.001 | 0.0969 | 0.1113 | 0.1317 | 0.1502 | 0.0032 | 0.0046 | 0.0063 | 0.0087 |
| 0.0001 | 0.0525 | 0.0664 | 0.0800 | 0.0815 | 0.0009 | 0.0014 | 0.0021 | 0.0029 |

Table C.26: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the unbiased MBA simulations, found using unbiased perturbed distances emulating the power transform (5.22) with non-metric MDS.

# D Bias correction simulation results

Appendix containing corrected model-based approach (MBA) simulation results, to compliment results in Section 5.6.1. All the results here use the power transform (4.6) and metric multidimensional scaling.

| $\theta_{1:p}$ | | | | $m_{min} = 1$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 27.6641 | 20.1971 | 13.5770 | 9.2932 | 30.1163 | 23.3146 | 17.6761 | 13.0922 |
| 0.01 | 57.5843 | 48.8636 | 40.0941 | 31.7826 | 58.6994 | 50.2041 | 41.6320 | 33.4073 |
| 0.001 | 81.2340 | 75.3676 | 68.3693 | 60.4333 | 81.8874 | 76.1971 | 69.4243 | 61.5050 |
| 0.0001 | 93.2328 | 90.6692 | 87.2824 | 82.8996 | 93.4521 | 91.0162 | 87.7429 | 83.4862 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.9544 | 18.2244 | 12.9223 | 9.8966 | 21.7780 | 14.7522 | 10.2674 | 8.5787 |
| 0.01 | 54.6915 | 45.7856 | 37.2079 | 29.1054 | 52.1277 | 43.3316 | 34.6282 | 26.7121 |
| 0.001 | 79.3632 | 73.1630 | 65.7557 | 57.6323 | 77.7610 | 71.1939 | 63.4900 | 55.1194 |
| 0.0001 | 92.4191 | 89.6242 | 85.8852 | 81.1692 | 91.7106 | 88.6561 | 84.6956 | 79.6223 |
| $\theta_{1:p}$ | | | | $m_{min} = 2$ | | | | |
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 27.8449 | 20.4622 | 14.5415 | 10.5089 | 30.0553 | 23.2310 | 17.5600 | 13.2833 |
| 0.01 | 57.5867 | 48.8794 | 40.0921 | 31.7644 | 58.6948 | 50.1146 | 41.6229 | 33.4613 |
| 0.001 | 81.2009 | 75.4126 | 68.4244 | 60.4337 | 81.9173 | 76.1680 | 69.2851 | 61.5411 |
| 0.0001 | 93.2342 | 90.6892 | 87.2926 | 82.9007 | 93.4945 | 90.9954 | 87.7320 | 83.5119 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 24.8992 | 18.3351 | 13.8181 | 11.1620 | 21.9026 | 15.9170 | 12.2934 | 10.5900 |
| 0.01 | 54.6804 | 45.8860 | 37.1511 | 29.0405 | 52.1268 | 43.3115 | 34.7174 | 26.7251 |
| 0.001 | 79.3781 | 73.1536 | 65.8700 | 57.5531 | 77.7960 | 71.2022 | 63.5694 | 55.1047 |
| 0.0001 | 92.4390 | 89.5965 | 85.9070 | 81.1780 | 91.7179 | 88.6770 | 84.6779 | 79.6025 |

Table D.27: Percentage of information projected into the first $k$ dimensions $\theta_{1:p}$ (2.11) from the bias corrected MBA simulations, found using bias corrected perturbed distances (5.27) for the power transform (4.6) with metric MDS. For the straight line $p = 1$ and for the parabola; semi-circle and circle $p = 2$. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | | | | $m_{min} = 1$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0590 | 0.0906 | 0.1314 | 0.1726 | 0.2913 | 0.4087 | 0.5114 | 0.5858 |
| 0.01 | 0.0156 | 0.0226 | 0.0320 | 0.0456 | 0.0631 | 0.0920 | 0.1349 | 0.2317 |
| 0.001 | 0.0050 | 0.0070 | 0.0097 | 0.0141 | 0.0191 | 0.0279 | 0.0387 | 0.0565 |
| 0.0001 | 0.0016 | 0.0022 | 0.0031 | 0.0044 | 0.0060 | 0.0086 | 0.0122 | 0.0171 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.1448 | 0.2906 | 0.4988 | 0.5343 | 0.0895 | 0.1475 | 0.1933 | 0.1848 |
| 0.01 | 0.0371 | 0.0541 | 0.0781 | 0.1109 | 0.0221 | 0.0315 | 0.0453 | 0.0663 |
| 0.001 | 0.0116 | 0.0166 | 0.0231 | 0.0328 | 0.0070 | 0.0098 | 0.0138 | 0.0197 |
| 0.0001 | 0.0037 | 0.0052 | 0.0074 | 0.0106 | 0.0022 | 0.0031 | 0.0044 | 0.0062 |

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | | | | $m_{min} = 2$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0566 | 0.0877 | 0.1134 | 0.1233 | 0.2906 | 0.3960 | 0.4664 | 0.5107 |
| 0.01 | 0.0156 | 0.0223 | 0.0314 | 0.0450 | 0.0647 | 0.0932 | 0.1411 | 0.2218 |
| 0.001 | 0.0049 | 0.0070 | 0.0096 | 0.0136 | 0.0190 | 0.0272 | 0.0386 | 0.0546 |
| 0.0001 | 0.0015 | 0.0022 | 0.0031 | 0.0044 | 0.0060 | 0.0085 | 0.0120 | 0.0171 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.1441 | 0.2169 | 0.3082 | 0.3198 | 0.0853 | 0.1231 | 0.1474 | 0.1388 |
| 0.01 | 0.0368 | 0.0525 | 0.0770 | 0.1133 | 0.0223 | 0.0317 | 0.0451 | 0.0656 |
| 0.001 | 0.0116 | 0.0165 | 0.0231 | 0.0336 | 0.0069 | 0.0099 | 0.0139 | 0.0196 |
| 0.0001 | 0.0037 | 0.0053 | 0.0074 | 0.0104 | 0.0022 | 0.0031 | 0.0044 | 0.0062 |

Table D.28: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the bias corrected MBA simulations, found using bias corrected perturbed distances (5.27) for the power transform (4.6) with metric MDS. Top table: $m_{\text{min}} = 1$ adjustment. Bottom table: $m_{\text{min}} = 2$ adjustment.

$G(\mathbf{X}, \hat{\mathbf{X}})$ $m_{min} = 1$

| | Straight line | | | | Parabola | | | |
|---|---|---|---|---|---|---|---|---|
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0111 | 1.0283 | 1.0721 | 1.1701 | 1.0123 | 1.0431 | 1.1127 | 1.2455 |
| 0.01 | 1.0005 | 1.0019 | 1.0033 | 1.0072 | 1.0019 | 1.0036 | 1.0044 | 1.0059 |
| 0.001 | 1.0001 | 1.0001 | 1.0001 | 1.0008 | 1.0002 | 1.0004 | 1.0008 | 1.0016 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 | 1.0002 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0130 | 1.0279 | 1.1240 | 1.2487 | 1.0220 | 1.0806 | 1.1897 | 1.3101 |
| 0.01 | 1.0013 | 1.0025 | 1.0060 | 1.0134 | 1.0016 | 1.0030 | 1.0072 | 1.0168 |
| 0.001 | 1.0001 | 1.0003 | 1.0005 | 1.0011 | 1.0001 | 1.0002 | 1.0005 | 1.0011 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 | 1.0000 | 1.0001 | 1.0001 | 1.0001 |

$G(\mathbf{X}, \hat{\mathbf{X}})$ $m_{min} = 2$

| | Straight line | | | | Parabola | | | |
|---|---|---|---|---|---|---|---|---|
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0107 | 1.0272 | 1.0604 | 1.1117 | 1.0041 | 1.0248 | 1.0666 | 1.1187 |
| 0.01 | 1.0004 | 1.0015 | 1.0032 | 1.0066 | 1.0020 | 1.0035 | 1.0035 | 1.0064 |
| 0.001 | 1.0001 | 1.0002 | 1.0003 | 1.0005 | 1.0002 | 1.0005 | 1.0008 | 1.0017 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0001 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 1.0159 | 1.0225 | 1.0657 | 1.1017 | 1.0218 | 1.0684 | 1.1221 | 1.1161 |
| 0.01 | 1.0013 | 1.0028 | 1.0059 | 1.0129 | 1.0013 | 1.0034 | 1.0073 | 1.0173 |
| 0.001 | 1.0001 | 1.0002 | 1.0005 | 1.0011 | 1.0001 | 1.0004 | 1.0005 | 1.0012 |
| 0.0001 | 1.0000 | 1.0000 | 1.0001 | 1.0000 | 1.0000 | 1.0000 | 1.0001 | 1.0001 |

Table D.29: Size expansion values $G(\mathbf{X}, \hat{\mathbf{X}})$ (5.6) from the bias corrected MBA simulations, found using bias corrected perturbed distances (5.27) for the power transform (4.6) with metric MDS. Top table: $m_{\min} = 1$ adjustment. Bottom table: $m_{\min} = 2$ adjustment.

# E Post-processing simulation results

Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) and variance score values $w(\tilde{\mathbf{D}})$ (6.7), from post-processing the MBA fitted configurations (or bias corrected fitted configurations) with smoothing splines (Section 6.4).

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0157 | 0.0629 | 0.0999 | 0.1301 | 0.1007 | 0.0739 | 0.2581 | 0.2622 |
| 0.01 | 0.0103 | 0.0135 | 0.0255 | 0.0452 | 0.0325 | 0.0465 | 0.0668 | 0.0757 |
| 0.001 | 0.0031 | 0.0049 | 0.0071 | 0.0095 | 0.0119 | 0.0153 | 0.0252 | 0.0289 |
| 0.0001 | 0.0010 | 0.0014 | 0.0018 | 0.0029 | 0.0038 | 0.0052 | 0.0069 | 0.0104 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0696 | 0.0715 | 0.1629 | 0.4331 | 0.0322 | 0.0881 | 0.1696 | 0.2819 |
| 0.01 | 0.0202 | 0.0279 | 0.0378 | 0.0553 | 0.0087 | 0.0145 | 0.0241 | 0.0471 |
| 0.001 | 0.0071 | 0.0098 | 0.0132 | 0.0180 | 0.0032 | 0.0043 | 0.0060 | 0.0085 |
| 0.0001 | 0.0023 | 0.0032 | 0.0044 | 0.0064 | 0.0010 | 0.0015 | 0.0021 | 0.0028 |

Table E.30: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the post-processing simulations for the exponential transform (4.4). The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original configuration $\mathbf{X}$ and the smoothed fitted configurations $\hat{\hat{\mathbf{X}}}$.

| $P(\mathbf{X}, \hat{\mathbf{X}})$ | **Power transform $m_{\text{min}} = 2$** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0047 | 0.0068 | 0.0053 | 0.0234 | 0.1736 | 0.2120 | 0.2137 | 0.3130 |
| 0.01 | 0.004 | 0.0046 | 0.0064 | 0.006 | 0.0155 | 0.0211 | 0.0447 | 0.1246 |
| 0.001 | 0.0014 | 0.0019 | 0.0029 | 0.0041 | 0.0068 | 0.0085 | 0.0129 | 0.0168 |
| 0.0001 | 0.0006 | 0.0007 | 0.0009 | 0.0015 | 0.0025 | 0.0036 | 0.0043 | 0.0067 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.0413 | 0.0635 | 0.0855 | 0.2437 | 0.0424 | 0.0552 | 0.0526 | 0.0602 |
| 0.01 | 0.0126 | 0.0173 | 0.0223 | 0.0309 | 0.0121 | 0.0170 | 0.0244 | 0.0352 |
| 0.001 | 0.0052 | 0.0063 | 0.0096 | 0.0126 | 0.0038 | 0.0051 | 0.0075 | 0.0105 |
| 0.0001 | 0.0019 | 0.0023 | 0.0034 | 0.0046 | 0.0014 | 0.0019 | 0.0025 | 0.0036 |

Table E.31: Shape difference values $P(\mathbf{X}, \hat{\mathbf{X}})$ (5.5) from the post-processing simulations for the power transform (4.4) with the $m_{\text{min}} = 2$ adjustment. The $P(\mathbf{X}, \hat{\mathbf{X}})$ values are found between the original configuration $\mathbf{X}$ and the smoothed fitted configurations $\hat{\hat{\mathbf{X}}}$.

| $w(\tilde{\mathbf{D}})$ | **Exponential transform** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 2.9119 | 2.1623 | 11.7841 | 28.3814 | 0.9946 | 0.2178 | 4.0275 | 20.1807 |
| 0.01 | 1.3038 | 1.5448 | 1.6885 | 3.3319 | 0.8170 | 0.4898 | 0.2145 | 0.1714 |
| 0.001 | 1.3227 | 1.5847 | 1.2776 | 1.3249 | 0.7015 | 0.8372 | 0.4469 | 0.5011 |
| 0.0001 | 1.1888 | 1.2287 | 1.541 | 1.1775 | 0.6890 | 0.8589 | 0.9734 | 0.6406 |
| | Semi-circle | | | | Circle | | | |
| $\alpha$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 0.7450 | 0.3391 | 0.5092 | 0.3611 | 0.4029 | 0.7547 | 1.0511 | 1.4260 |
| 0.01 | 0.7590 | 0.4739 | 0.2306 | 0.3024 | 0.7978 | 0.5209 | 0.3016 | 0.5153 |
| 0.001 | 0.7391 | 0.7320 | 0.4245 | 0.5129 | 0.6904 | 0.6531 | 0.4582 | 0.3696 |
| 0.0001 | 0.6508 | 0.8521 | 0.9660 | 0.7465 | 0.6431 | 0.7431 | 0.7049 | 0.8220 |

Table E.32: Variance score values $w(\tilde{\mathbf{D}})$ (6.7) from the post-processing simulations for the exponential transform (4.4).

| $w(\tilde{\mathbf{D}})$ | **Power transform** $m_{\text{min}} = 2$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Straight line | | | | Parabola | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 15.6367 | 23.1083 | 34.0986 | 53.5346 | 25.9695 | 30.7442 | 41.6738 | 60.2623 |
| 0.01 | 2.8691 | 3.5594 | 2.749 | 6.1202 | 3.0600 | 3.5256 | 6.8311 | 16.4996 |
| 0.001 | 2.0324 | 1.9586 | 2.0247 | 2.0499 | 1.4956 | 2.0005 | 2.0119 | 1.8551 |
| 0.0001 | 1.8199 | 1.4785 | 1.9867 | 1.8530 | 1.9266 | 1.8844 | 1.5389 | 1.4694 |
| | Semi-circle | | | | Circle | | | |
| $b_0$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ | $\rho = 1$ | $\rho = 2$ | $\rho = 4$ | $\rho = 8$ |
| 0.1 | 21.4561 | 30.2564 | 41.3704 | 56.379 | 24.0485 | 31.6124 | 42.9289 | 67.0777 |
| 0.01 | 3.3914 | 3.6165 | 4.8131 | 8.9179 | 1.3958 | 2.6556 | 4.5349 | 9.3891 |
| 0.001 | 1.4891 | 1.7797 | 1.6724 | 2.6190 | 0.7925 | 1.1133 | 0.9522 | 1.3913 |
| 0.0001 | 1.4951 | 2.0426 | 1.9536 | 1.2949 | 0.5971 | 0.6042 | 0.8934 | 0.6262 |

Table E.33: Variance score values $w(\tilde{\mathbf{D}})$ (6.7) from the post-processing simulations for the power transform (4.6) with the $m_{\text{min}} = 2$ adjustment.

# F Parameter estimation results

## F.1 Estimating $\alpha$ for the exponential transform

Table of $\alpha$ estimates for dispersion $\rho = 1$ can be found in Table 6.3.

| $\rho = 2$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $\alpha$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ |
| 0.1 | 0.082039 | 17.9611 | 0.085209 | 14.7915 |
| 0.01 | 0.009730 | 2.7039 | 0.009693 | 3.0732 |
| 0.001 | 0.001001 | 0.1187 | 0.000982 | 1.768 |
| 0.0001 | 0.000090 | 9.5576 | 0.000090 | 9.8301 |

Table F.34: $\alpha$ estimates $(\hat{\alpha})$ from the parameter estimation simulations for the exponential transform (4.4). The $\hat{\alpha}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle, using the exponential transform and with dispersion $\rho = 2$. Column one gives the different levels of $\alpha$ used. For each score function the mean $\hat{\alpha}$ and the mean percentage error between $\hat{\alpha}$ and $\alpha$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 4$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $\alpha$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ |
| 0.1 | 0.074034 | 25.9664 | 0.084336 | 15.6635 |
| 0.01 | 0.009468 | 5.322 | 0.00946 | 5.4037 |
| 0.001 | 0.000993 | 0.666 | 0.001004 | 0.3643 |
| 0.0001 | 0.000091 | 9.2358 | 0.00009 | 9.8301 |

Table F.35: $\alpha$ estimates $(\hat{\alpha})$ from the parameter estimation simulations for the exponential transform (4.4). The $\hat{\alpha}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle, using the exponential transform and with dispersion $\rho = 4$. Column one gives the different levels of $\alpha$ used. For each score function the mean $\hat{\alpha}$ and the mean percentage error between $\hat{\alpha}$ and $\alpha$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 8$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $\alpha$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ | $\hat{\alpha}$ | $\frac{\lvert\alpha-\hat{\alpha}\rvert}{\alpha}\%$ |
| 0.1 | 0.063328 | 36.6717 | 0.089217 | 10.783 |
| 0.01 | 0.009105 | 8.9452 | 0.009148 | 8.5188 |
| 0.001 | 0.000993 | 0.6503 | 0.000993 | 0.7081 |
| 0.0001 | 0.00009 | 9.609 | 0.00009 | 9.8301 |

Table F.36: $\alpha$ estimates $(\hat{\alpha})$ from the parameter estimation simulations for the exponential transform (4.4). The $\hat{\alpha}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle, using the exponential transform and with dispersion $\rho = 8$. Column one gives the different levels of $\alpha$ used. For each score function the mean $\hat{\alpha}$ and the mean percentage error between $\hat{\alpha}$ and $\alpha$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

## F.2 Estimating the $\beta$ for the power transform

**When** $\beta = -0.3$

| $\rho = 1$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.27032 | 9.8932 | -0.279115 | 6.9617 |
| 0.01 | -0.300124 | 0.0415 | -0.297763 | 0.7456 |
| 0.001 | -0.300482 | 0.1608 | -0.299788 | 0.0708 |
| 0.0001 | -0.30002 | 0.0065 | -0.299986 | 0.0046 |

Table F.37: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.3$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 1$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 2$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.253543 | 15.4858 | -0.262938 | 12.3541 |
| 0.01 | -0.295274 | 1.5754 | -0.295669 | 1.4438 |
| 0.001 | -0.301249 | 0.4163 | -0.299602 | 0.1328 |
| 0.0001 | -0.300034 | 0.0113 | -0.299952 | 0.0159 |

Table F.38: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.3$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 2$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 4$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.241904 | 19.3653 | -0.239686 | 20.1045 |
| 0.01 | -0.287838 | 4.054 | -0.292154 | 2.6154 |
| 0.001 | -0.302519 | 0.8396 | -0.299155 | 0.2818 |
| 0.0001 | -0.300065 | 0.0215 | -0.299891 | 0.0363 |

Table F.39: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.3$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 4$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 8$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.231561 | 22.8131 | -0.217662 | 27.446 |
| 0.01 | -0.277352 | 7.5493 | -0.284312 | 5.2294 |
| 0.001 | -0.301153 | 0.3845 | -0.298336 | 0.5547 |
| 0.0001 | -0.30031 | 0.1032 | -0.299827 | 0.0577 |

Table F.40: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.3$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 8$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

**When** $\beta = -0.5$

Table of $\beta$ estimates for $\beta = -0.5$ and dispersion $\rho = 1$ can be found in Table 6.4.

| $\rho = 2$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.382379 | 23.5242 | -0.369038 | 26.1924 |
| 0.01 | -0.470052 | 5.9895 | -0.478742 | 4.2516 |
| 0.001 | -0.499881 | 0.0237 | -0.497469 | 0.5063 |
| 0.0001 | -0.499976 | 0.0048 | -0.499788 | 0.0424 |

Table F.41: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.5$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 2$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 4$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.353238 | 29.3524 | -0.324164 | 35.1672 |
| 0.01 | -0.454445 | 9.1111 | -0.461102 | 7.7796 |
| 0.001 | -0.496053 | 0.7894 | -0.495197 | 0.9607 |
| 0.0001 | -0.500191 | 0.0381 | -0.499512 | 0.0976 |

Table F.42: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.5$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 4$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 8$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.333863 | 33.2274 | -0.288154 | 42.3693 |
| 0.01 | -0.431266 | 13.7469 | -0.434367 | 13.1265 |
| 0.001 | -0.48747 | 2.5059 | -0.491157 | 1.7686 |
| 0.0001 | -0.500645 | 0.129 | -0.499079 | 0.1841 |

Table F.43: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.5$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 8$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

**When** $\beta = -0.7$

| $\rho = 1$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.534021 | 23.7113 | -0.511856 | 26.8777 |
| 0.01 | -0.655224 | 6.3966 | -0.668485 | 4.5022 |
| 0.001 | -0.697656 | 0.3348 | -0.696595 | 0.4865 |
| 0.0001 | -0.69991 | 0.0128 | -0.699778 | 0.0317 |

Table F.44: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.7$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 1$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 2$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.484873 | 30.7324 | -0.44259 | 36.7728 |
| 0.01 | -0.631453 | 9.7924 | -0.644745 | 7.8936 |
| 0.001 | -0.690457 | 1.3633 | -0.693231 | 0.967 |
| 0.0001 | -0.699939 | 0.0087 | -0.699356 | 0.092 |

Table F.45: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.7$; with the $m_{\text{min}} = 2$ adjustment and with dispersion $\rho = 2$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

| $\rho = 4$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.446596 | 36.2006 | -0.383448 | 45.2217 |
| 0.01 | -0.606545 | 13.3508 | -0.608217 | 13.1119 |
| 0.001 | -0.678677 | 3.0462 | -0.686785 | 1.8879 |
| 0.0001 | -0.699856 | 0.0205 | -0.698732 | 0.1811 |

Table F.46: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.7$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 4$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

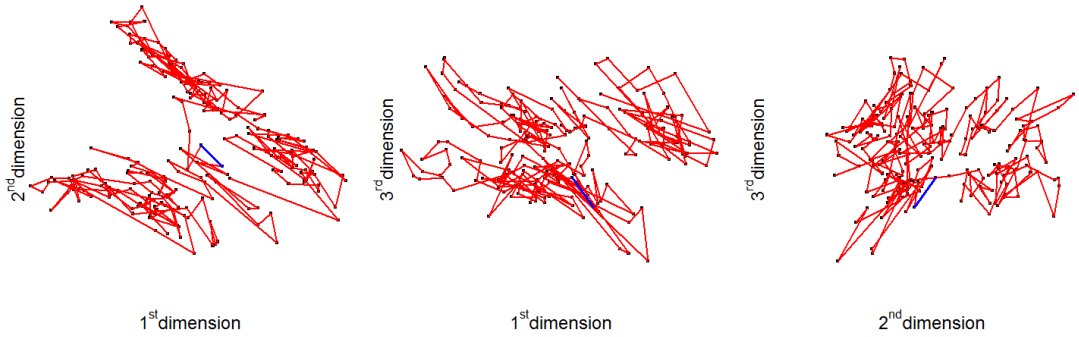
| $\rho = 8$ | $\chi^2$ | | $R_k(\hat{\mathbf{X}})$ | |
|---|---|---|---|---|
| $b_0$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ | $\hat{\beta}$ | $\frac{\lvert\beta-\hat{\beta}\rvert}{\beta}\%$ |
| 0.1 | -0.420946 | 39.8649 | -0.339711 | 51.4699 |
| 0.01 | -0.57111 | 18.4128 | -0.554672 | 20.7612 |
| 0.001 | -0.66113 | 5.5529 | -0.674504 | 3.6423 |
| 0.0001 | -0.698933 | 0.1524 | -0.697013 | 0.4267 |

Table F.47: $\beta$ estimates $(\hat{\beta})$ from the parameter estimation simulations for the power transform (4.6). The $\hat{\beta}$ values are found by applying the fitting algorithm (Section 4.1.3), with either the $\chi^2$ (4.9) or $R_k(\hat{\mathbf{X}})$ (5.24) score function and metric MDS, to the perturbed distance matrix $\tilde{\mathbf{D}}$. The matrix $\tilde{\mathbf{D}}$ is generated from a semi-circle using the power transform with $\beta = -0.7$; with the $m_{\min} = 2$ adjustment and with dispersion $\rho = 8$. Column one gives the different levels of $b_0$ used. For each score function the mean $\hat{\beta}$ and the mean percentage error between $\hat{\beta}$ and $\beta$ is given; in columns two and three for the $\chi^2$ score function, and in columns four and five for the $R_k(\hat{\mathbf{X}})$ score function.

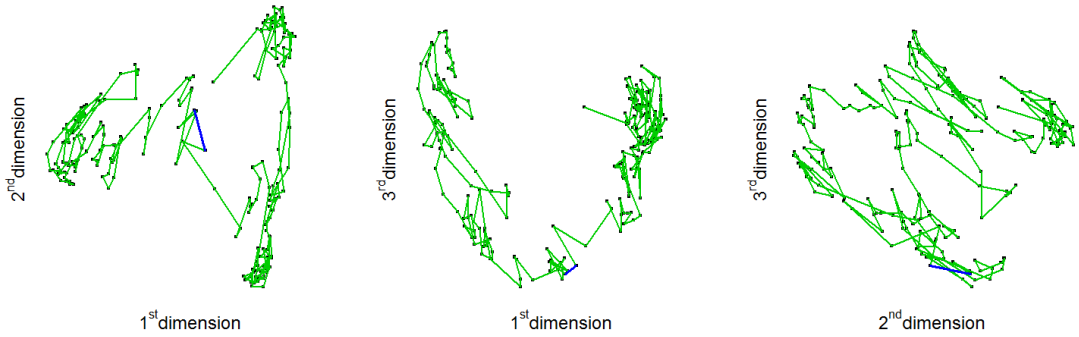
# G Estimated chromosome configuration

The four estimated chromosome configurations for each chromosome ($\hat{\mathbf{X}}_{E,M}, \hat{\mathbf{X}}_{P,M}, \hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{E,NM}$) can be sorted into two groups. The configurations within the groups sharing a similar shape, and comparing configurations in different groups shape should be different. Grouping the four estimated chromosome configuration is done using the process outlines in section 4.3.3.

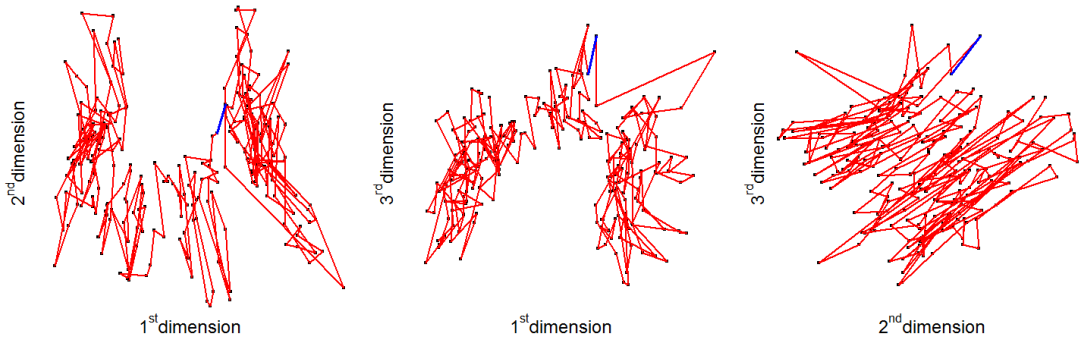
For each of the 22 chromosomes and the X chromosome an estimated configuration representative of each group is plotted, giving two estimated chromosome configurations plotted per chromosome. Group one is representative by the red configuration and group two representative by the green configuration. The blue line denotes the location of the centromere in the configuration; if no blue line is present then the centromere is found at the end of the chromosome and has been excluded for the estimated chromosome configuration. Table 4.6 summarizes how the estimated chromosome configurations are grouped.



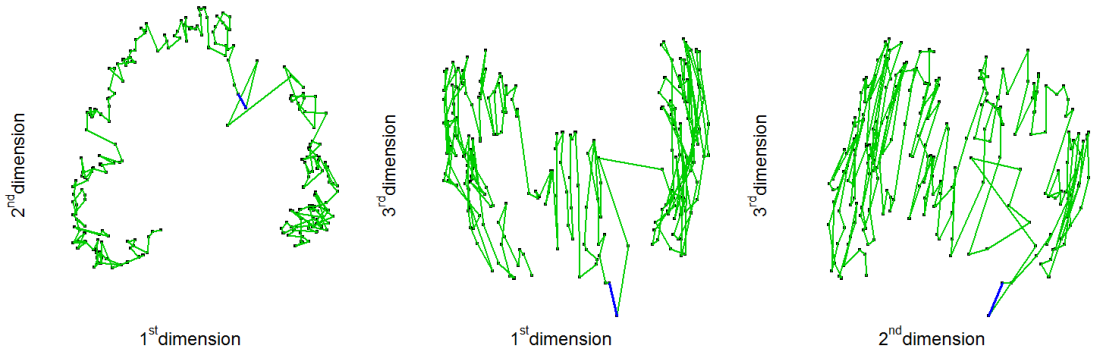

Chromosome 1 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



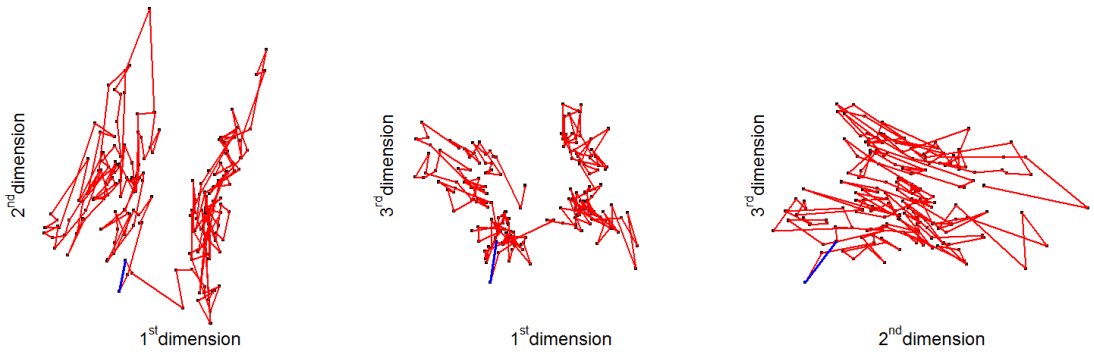

Chromosome 1 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



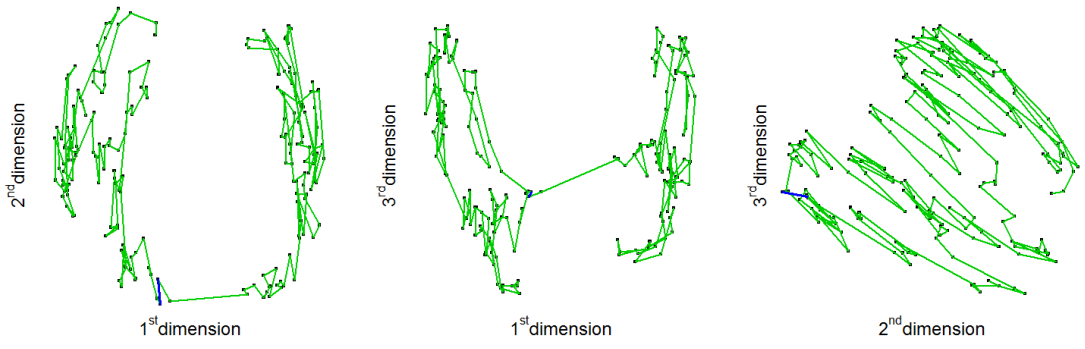

Chromosome 2 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



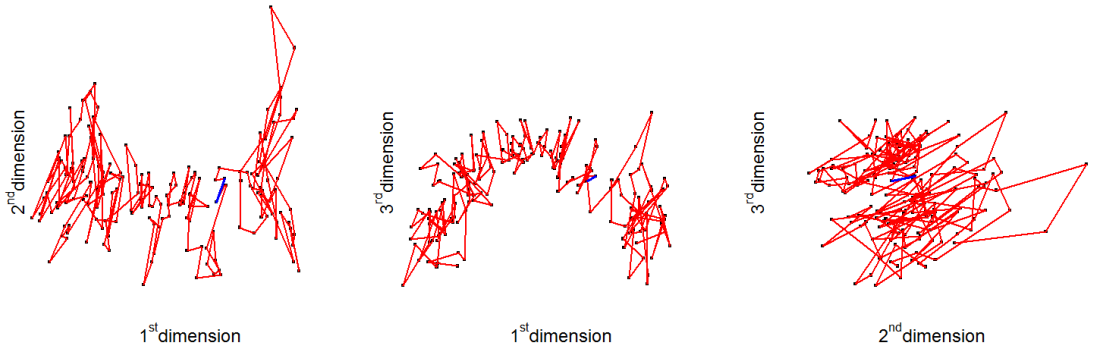

Chromosome 2 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



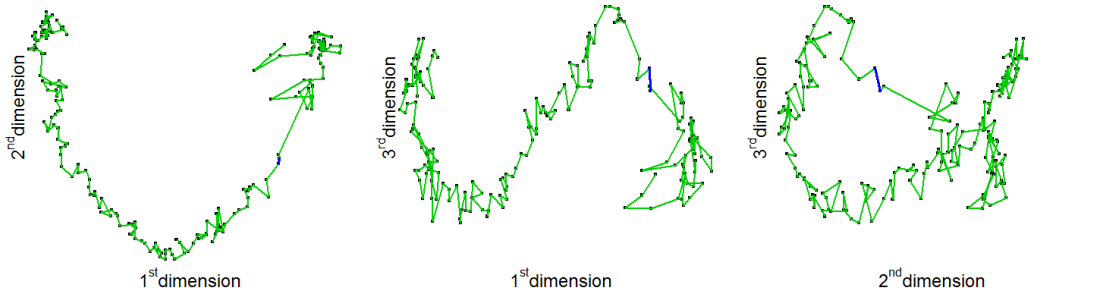

Chromosome 3 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



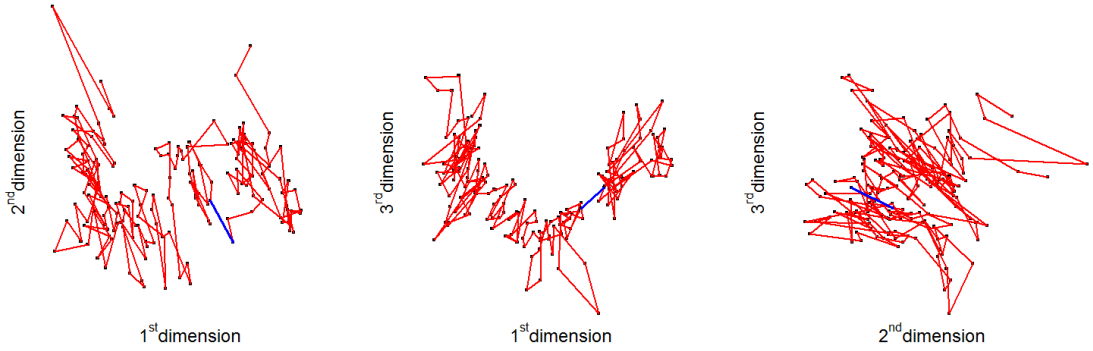

Chromosome 3 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



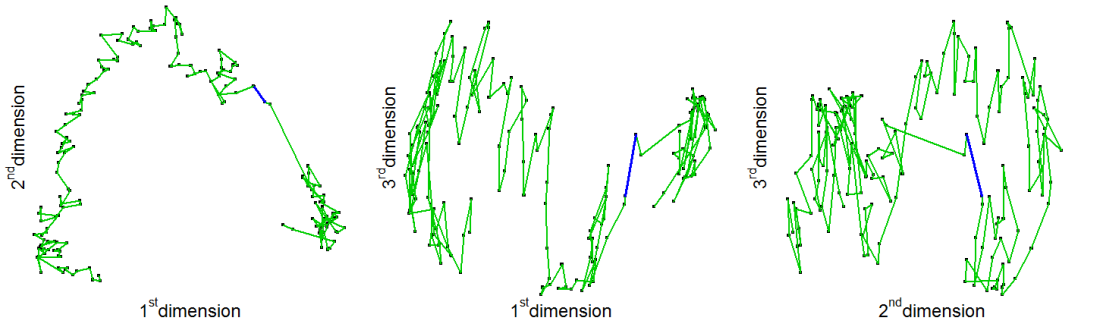

Chromosome 4 group one ($\hat{\mathbf{X}}_{P,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



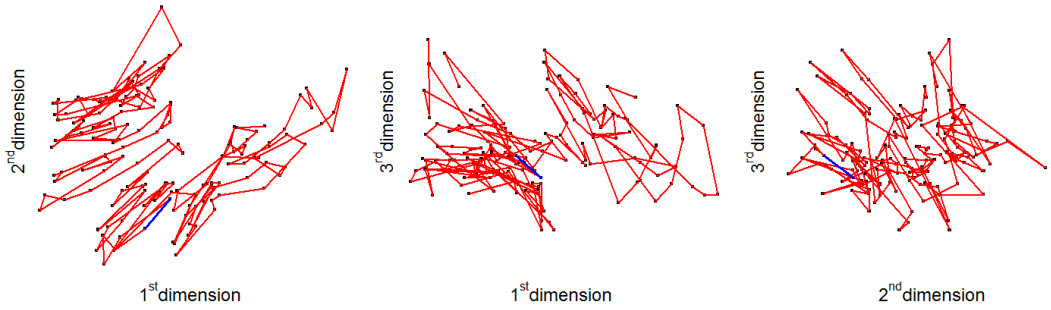

Chromosome 4 group two ($\hat{\mathbf{X}}_{E,M}$) estimated configuration.



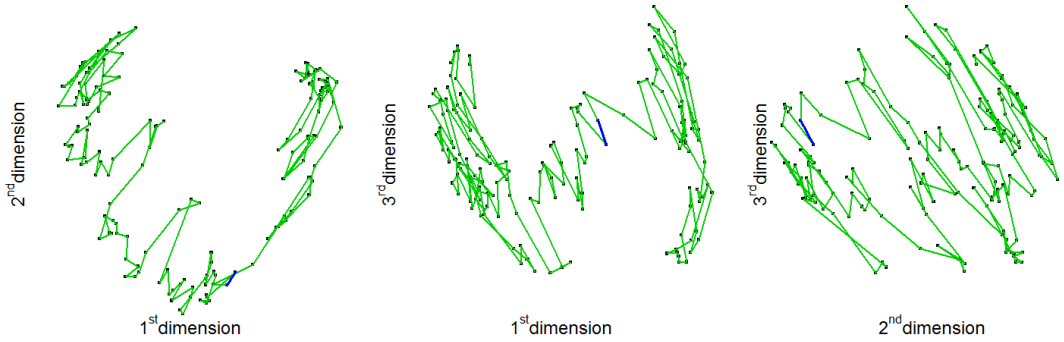

Chromosome 5 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



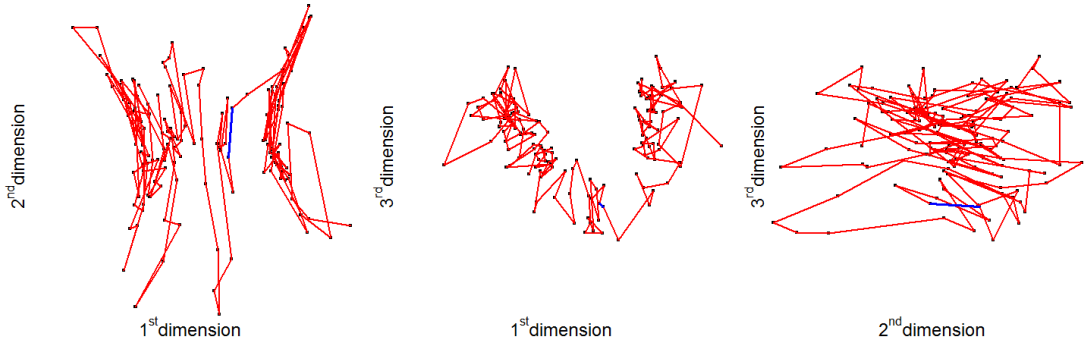

Chromosome 5 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



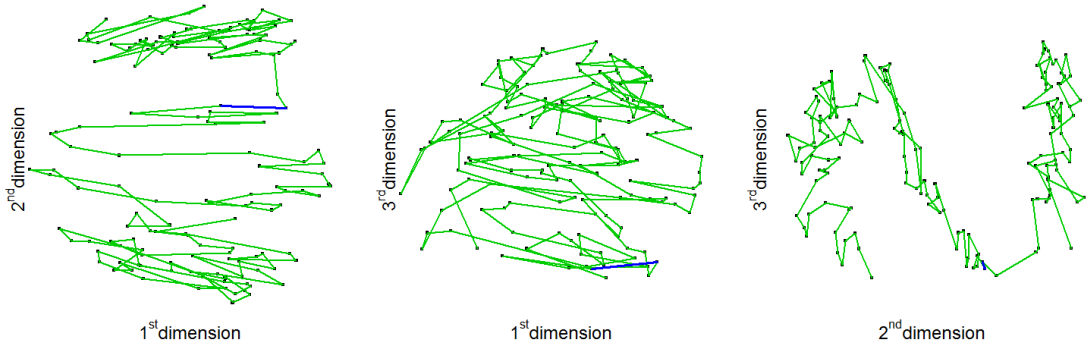

Chromosome 6 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



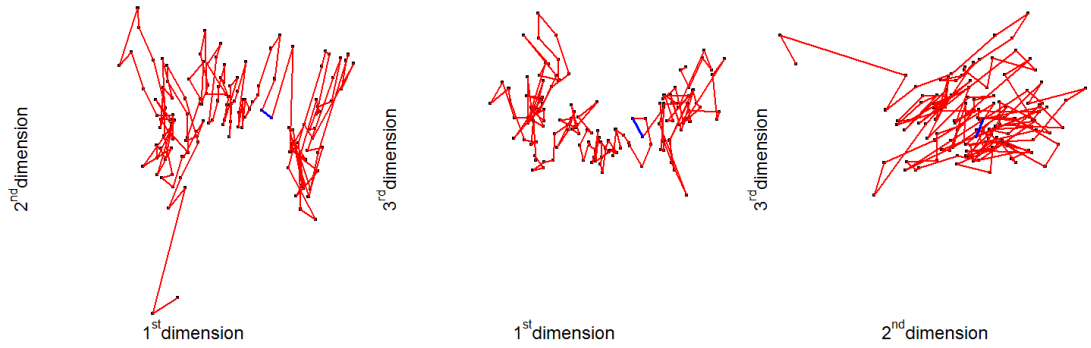

Chromosome 6 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



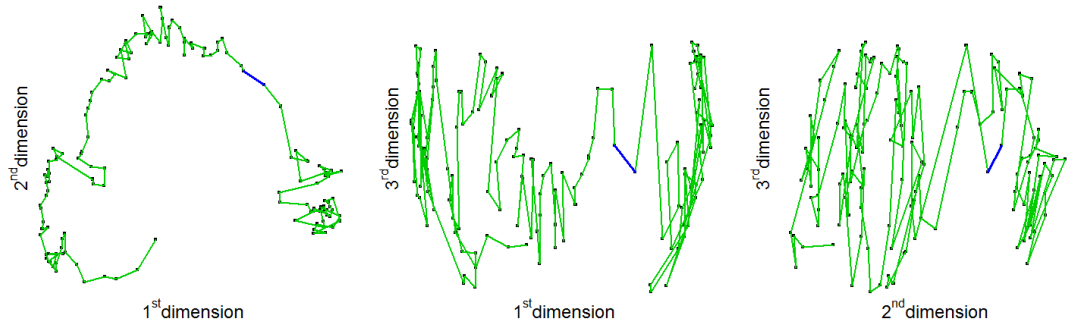

Chromosome 7 group one ($\hat{\mathbf{X}}_{P,M}$, $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



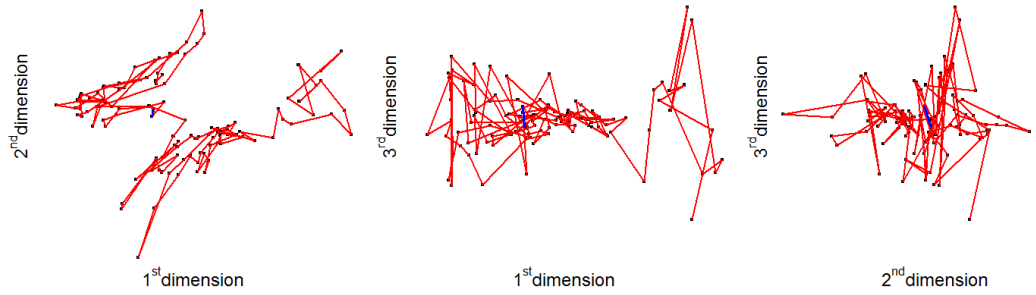

Chromosome 7 group two ($\hat{\mathbf{X}}_{E,NM}$) estimated configuration.



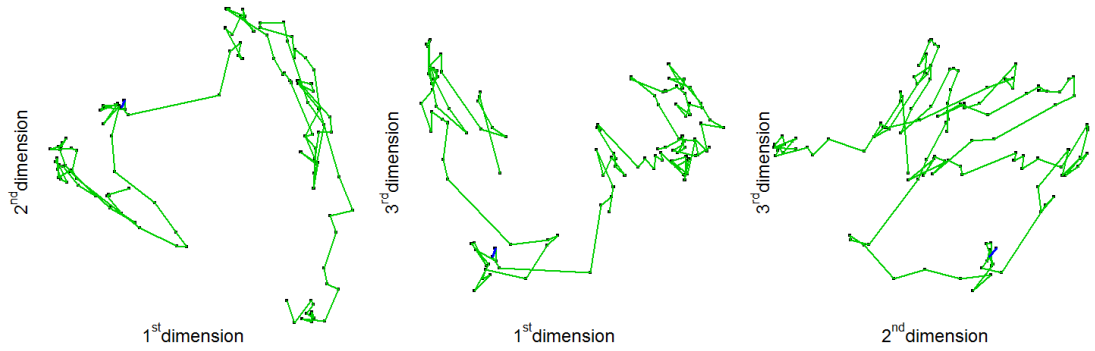

Chromosome 8 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



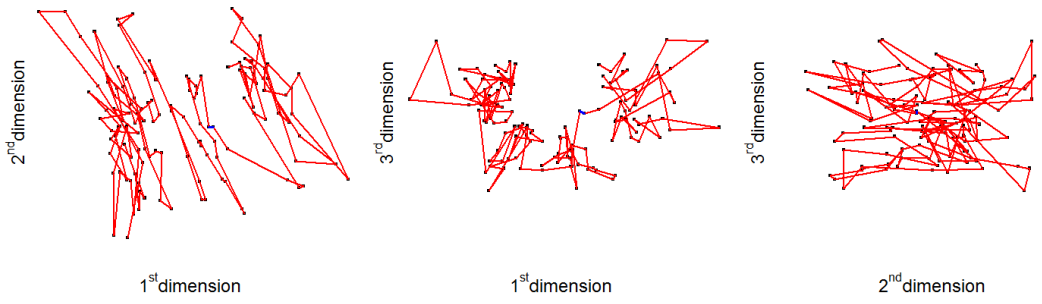

Chromosome 8 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



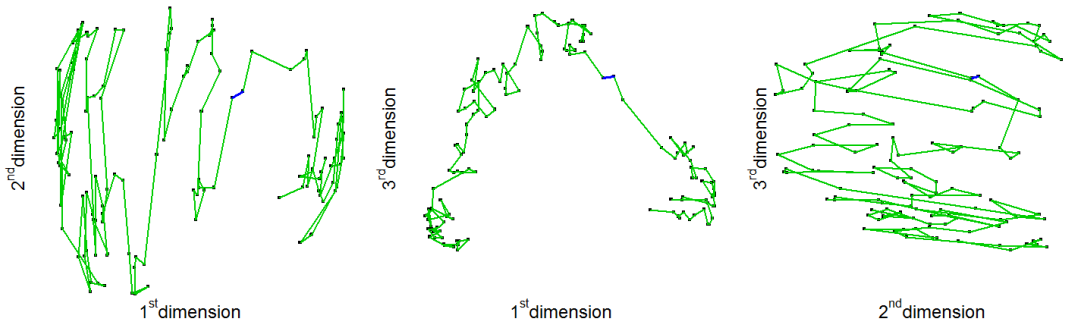

Chromosome 9 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



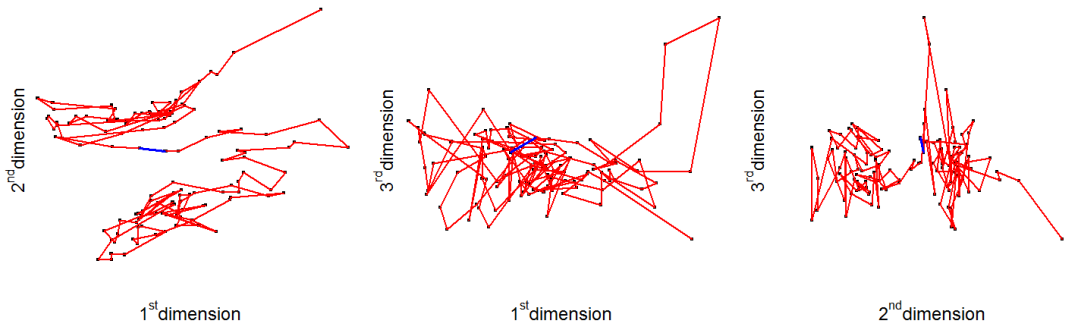

Chromosome 9 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



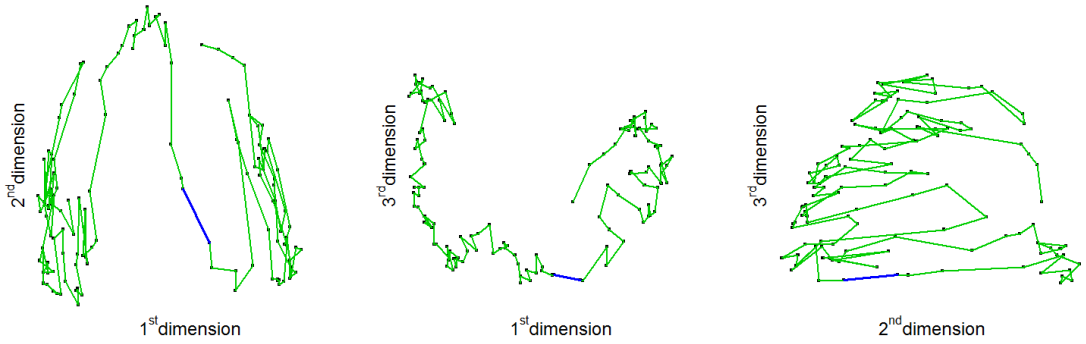

Chromosome 10 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



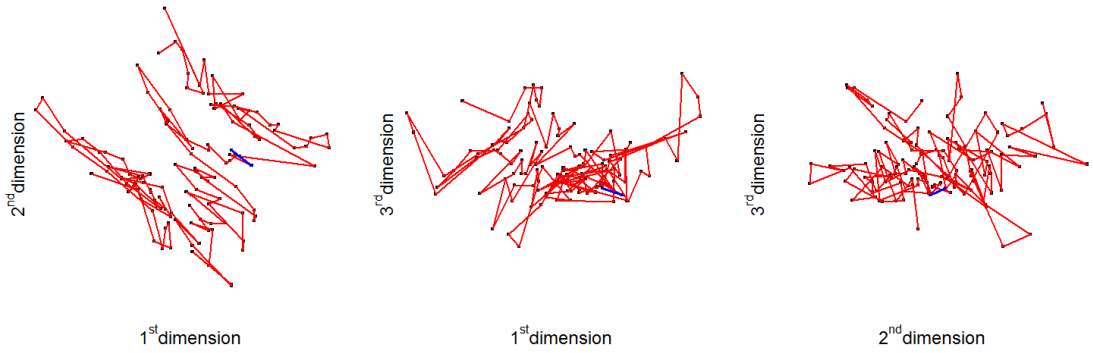

Chromosome 10 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



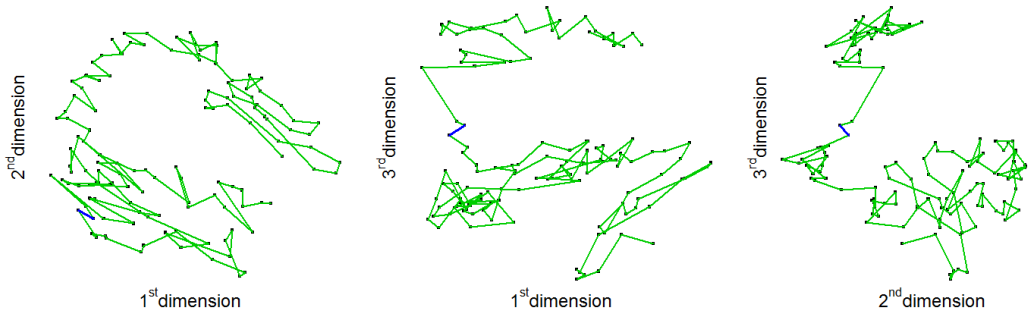

Chromosome 11 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



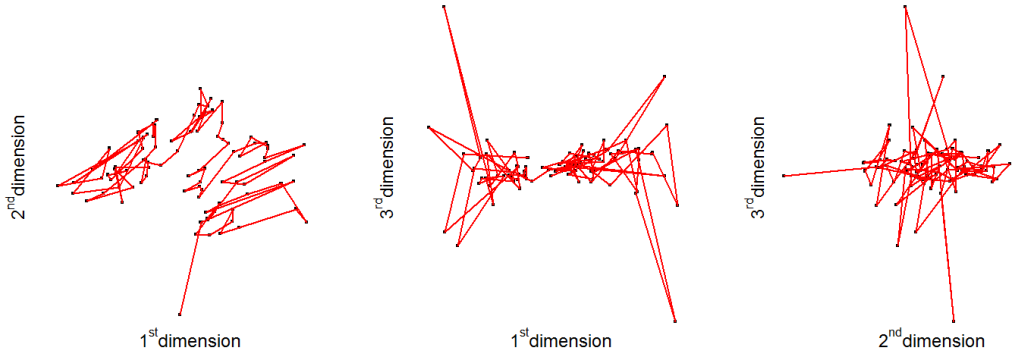

Chromosome 11 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



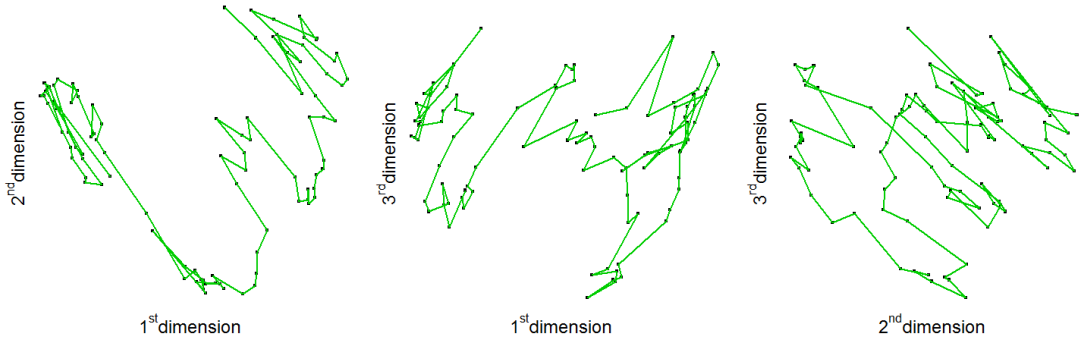

Chromosome 12 group one ($\hat{\mathbf{X}}_{P,M}$, $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



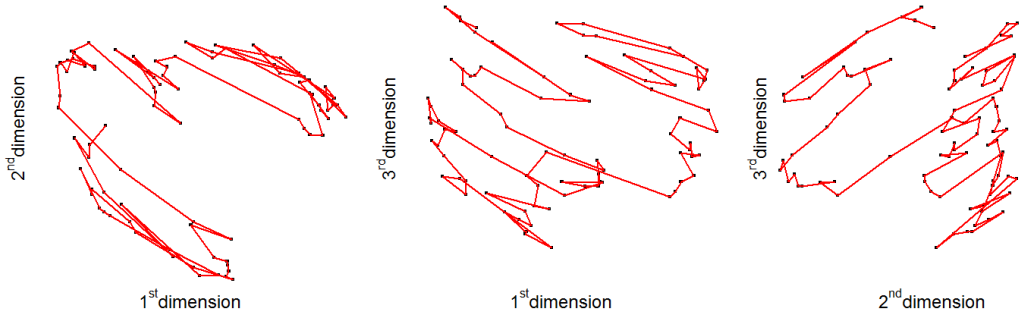

Chromosome 12 group two ($\hat{\mathbf{X}}_{E,NM}$) estimated configuration.



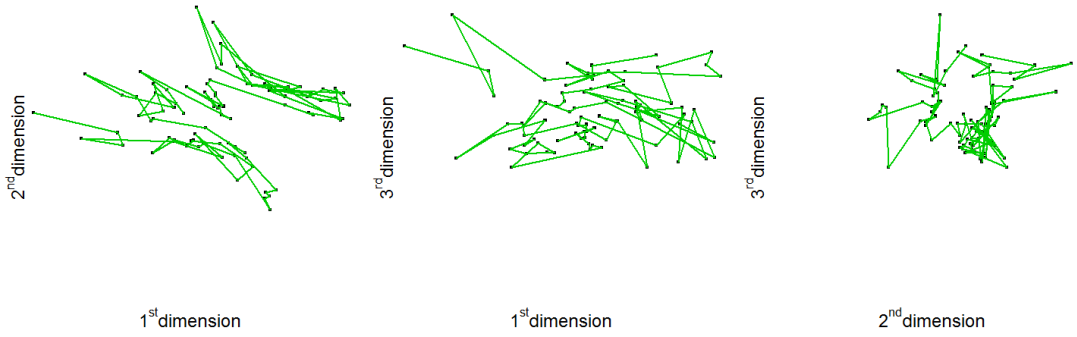

Chromosome 13 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



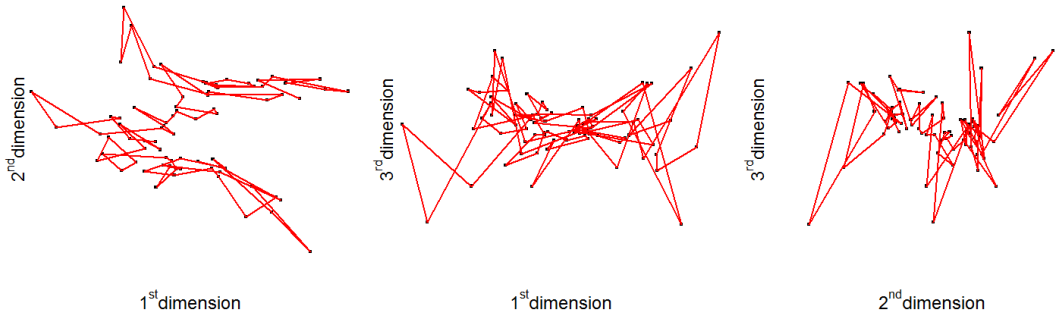

Chromosome 13 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



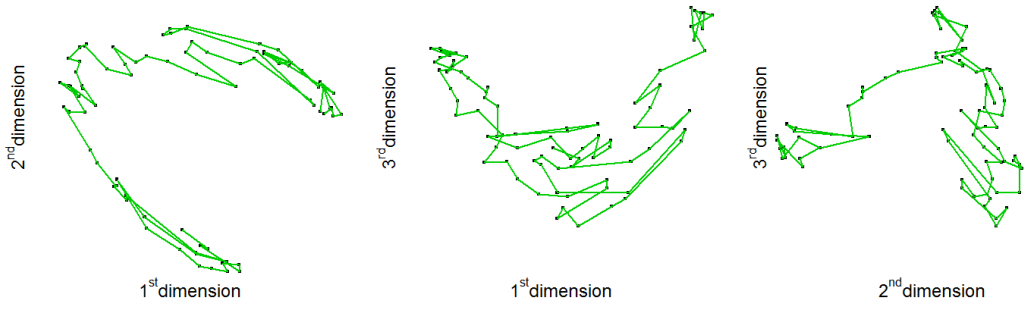

Chromosome 14 group one ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



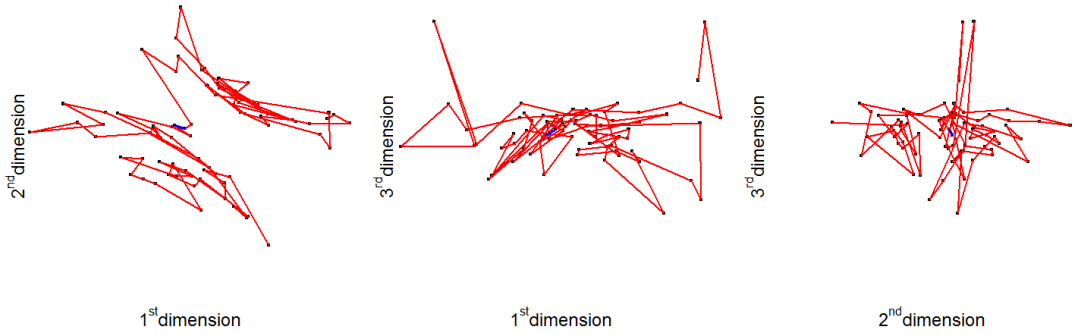

Chromosome 14 group two ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



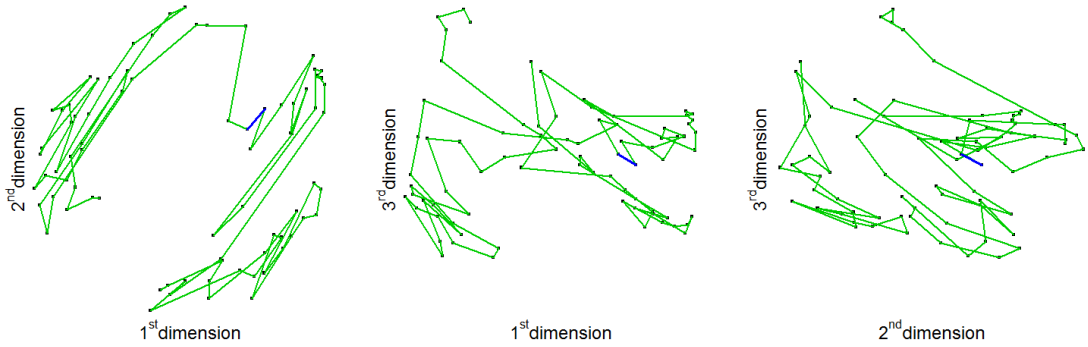

Chromosome 15 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



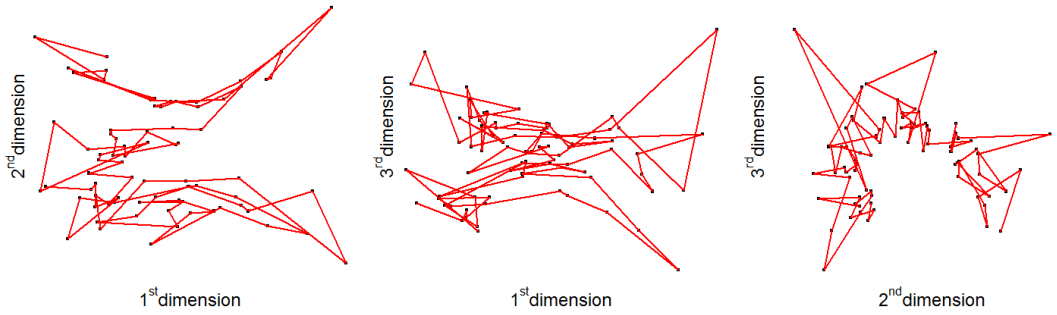

Chromosome 15 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



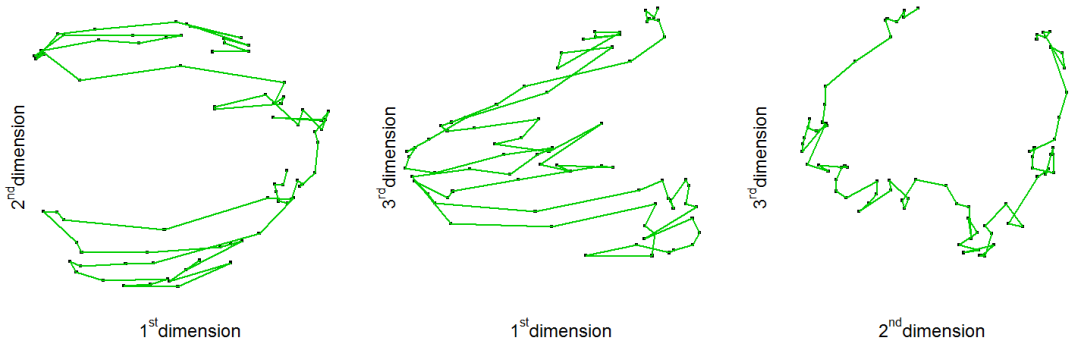

Chromosome 16 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



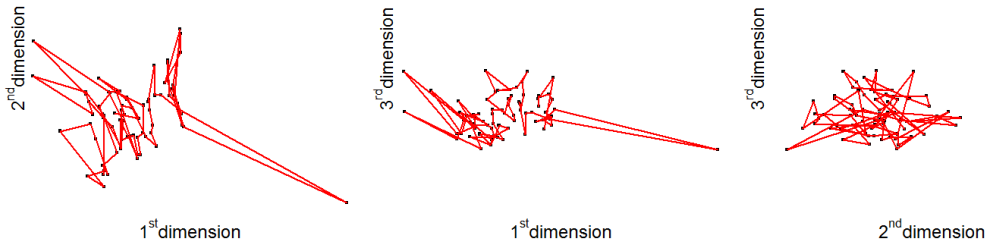

Chromosome 16 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



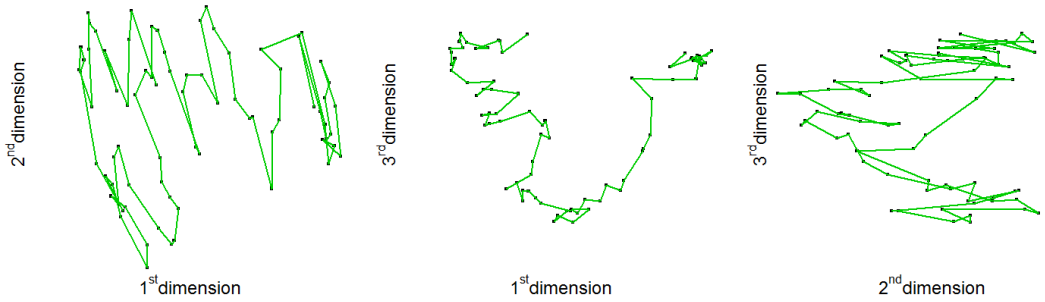

Chromosome 17 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



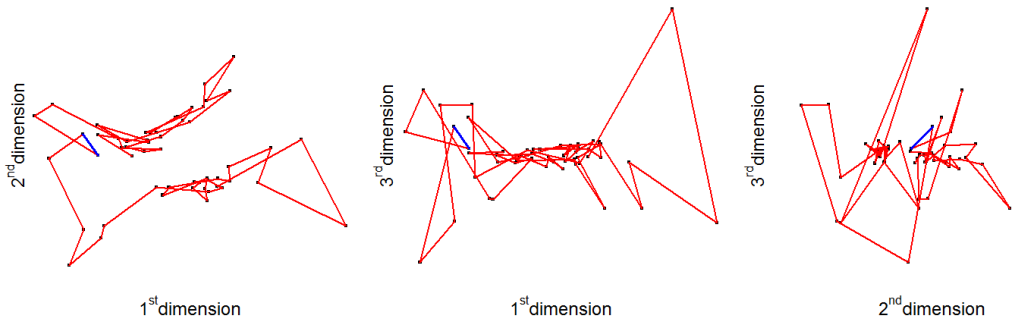

Chromosome 17 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



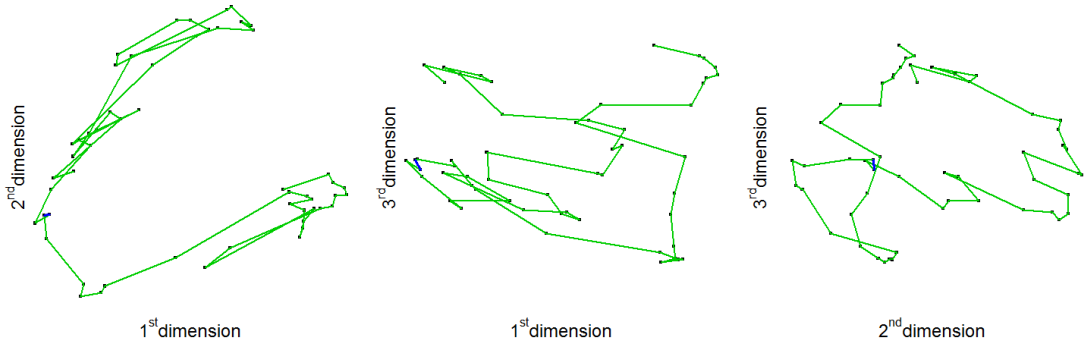

Chromosome 18 group one ($\hat{\mathbf{X}}_{P,M}$ and $\hat{\mathbf{X}}_{E,NM}$) estimated configuration.



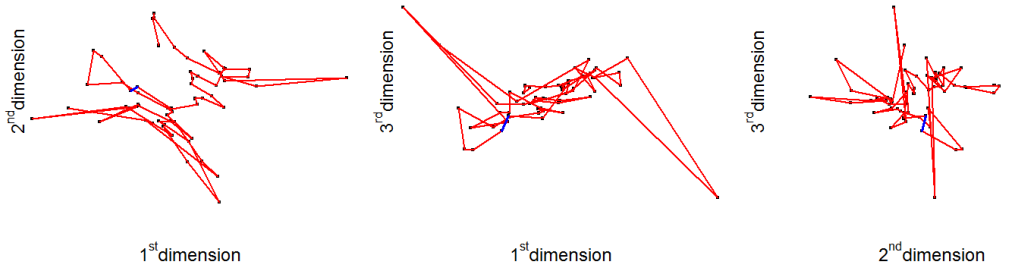

Chromosome 18 group two ($\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



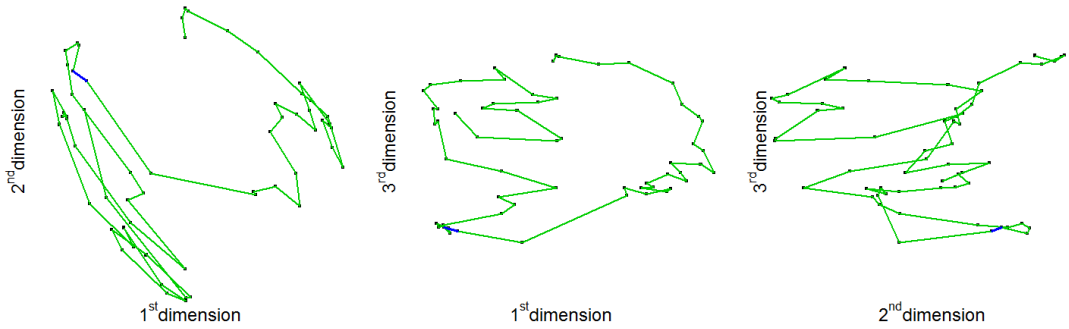

Chromosome 19 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



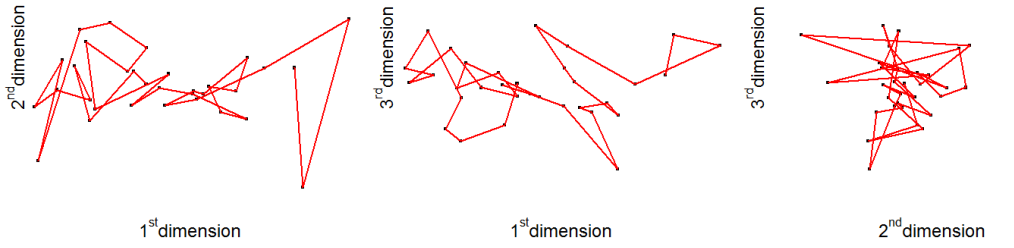

Chromosome 19 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



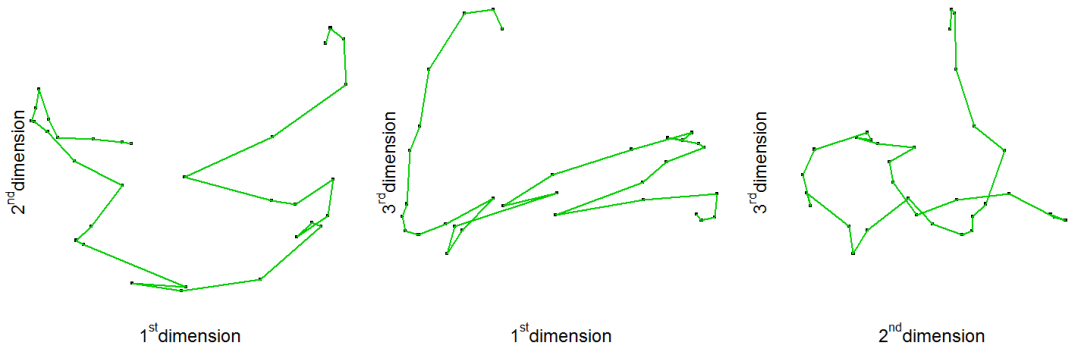

Chromosome 20 group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.



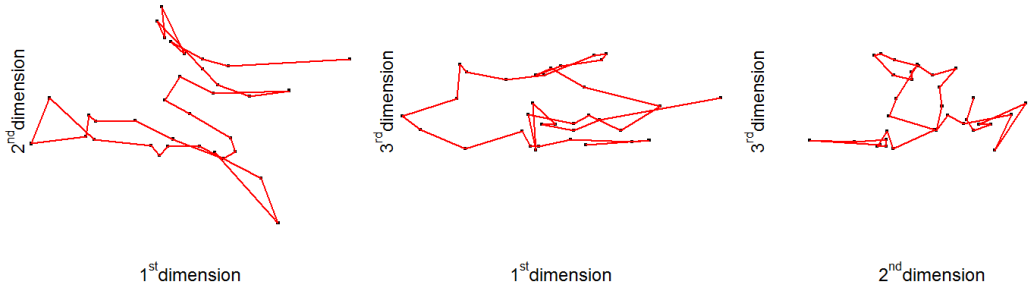

Chromosome 20 group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



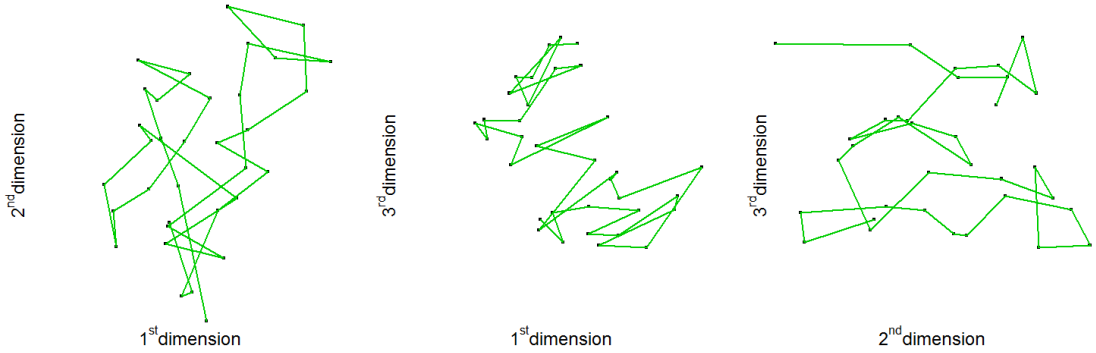

Chromosome 21 group one ($\hat{\mathbf{X}}_{P,M}$ and $\hat{\mathbf{X}}_{E,NM}$) estimated configuration.



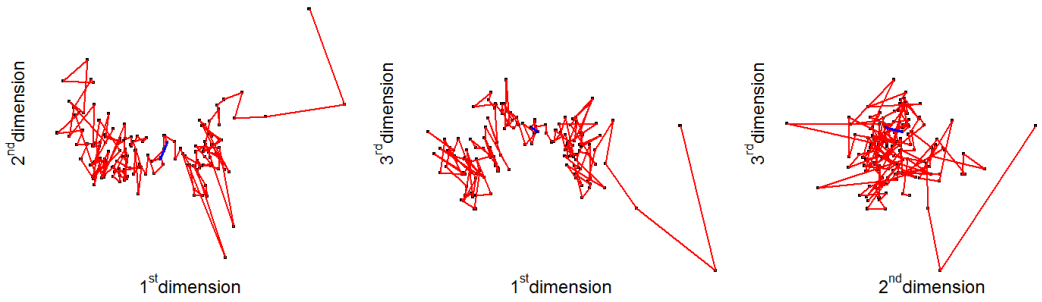

Chromosome 21 group two ($\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.



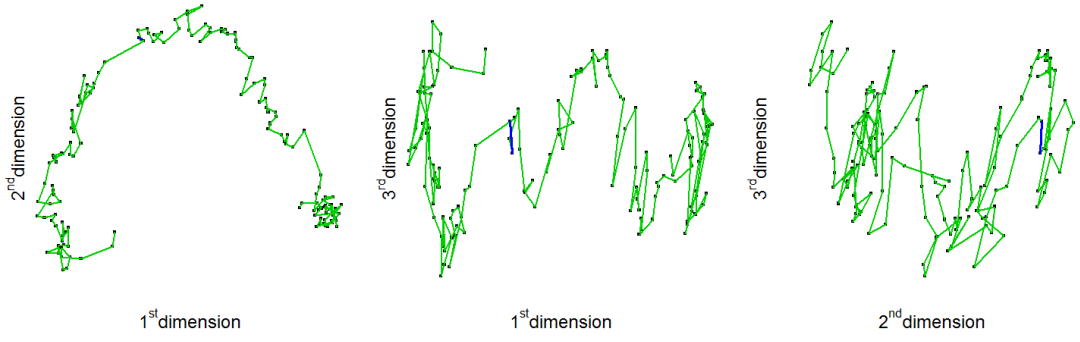

Chromosome 22 group one ($\hat{\mathbf{X}}_{P,M}$, $\hat{\mathbf{X}}_{E,M}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.




Chromosome 22 group two ($\hat{\mathbf{X}}_{E,NM}$) estimated configuration.




Chromosome X group one ($\hat{\mathbf{X}}_{P,M}$) estimated configuration.




Chromosome X group two ($\hat{\mathbf{X}}_{E,M}$, $\hat{\mathbf{X}}_{E,NM}$ and $\hat{\mathbf{X}}_{P,NM}$) estimated configuration.

# Bibliography

Akhtar, A. and S. M. Gasser (2007). The nuclear envelope and transcriptional control. *Nature Reviews Genetics 8*, 507–517.

Barbieri, M., M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, and M. Nicodemi (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences 109*, 16173–16178.

Baù, D., A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom (2011). The three-dimensional folding of the $\alpha$-globin gene domain reveals formation of chromatin globules. *Nature Structural & Molecular Biology 18*, 107–114.

Ben-Elazar, S., Z. Yakhini, and I. Yanai (2013). Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Research 41*, 2191–2201.

Bolzer, A., G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biology 3*, e157.

Boyle, S., S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics 10*, 211–219.

Buja, A. and D. F. Swayne (2002). Visualization methodology for multidimensional scaling. *Journal of Classification 19*, 7–43.

Christensen, R. R. (1990). *Log-Linear Models. Texts in Statistics*. Springer-Verlag, New York.

Cox, T. F. and M. A. Cox (2000). *Multidimensional Scaling*. CRC Press.

Cremer, T. and M. Cremer (2010). Chromosome territories. *Cold Spring Harbor perspectives in biology 2*, a003889.

Cremer, T., A. Kurz, R. Zirbel, S. Dietzel, B. Rinke, E. Schröck, M. R. Speicher, U. Mathieu, A. Jauch, P. Emmerich, H. Scherthan, T. Ried, C. Cremer, and P. Lichter (1993). The role of chromosome territories in the funtional compartmentalization of the cell nucleus. In *Cold Spring Harbor Symposia on Quantitative Biology*, pp. 777–792.

Croft, J. A., J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore (1999). Differences in the localization and morphology of chromosomes in the human nucleus. *The Journal of Cell Biology 145*, 1119–1131.

De Leeuw, J. (2008). A horseshoe for multidimensional scaling. *Department of Statistics, UCLA*.

De Leeuw, J. (2011). Applications of convex analysis to multidimensional scaling. *Department of Statistics, UCLA*.

de Wit, E. and W. de Laat (2012). A decade of 3c technologies: insights into nuclear organization. *Genes & Development 26*, 11–24.

Dekker, J., M. A. Marti-Renom, and L. A. Mirny (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics 14*, 390–403.

Dekker, J., K. Rippe, M. Dekker, and N. Kleckner (2002). Capturing chromosome conformation. *Science 295*, 1306–1311.

Diaconis, P., S. Goel, and S. Holmes (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, 777–807.

Dobson, A. J. and A. G. Barnett (2008). *An Introduction to Generalized Linear Models*. CRC Press.

Dostie, J., T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, et al. (2006). Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements. *Genome Research 16*, 1299–1309.

Duan, Z., M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble (2010). A three-dimensional model of the yeast genome. *Nature 465*, 363–367.

Everitt, B., S. Landau, and M. Leese (2001). *Cluster Analysis , London: Arnold*.

Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer.

Goetze, S., J. Mateos-Langerak, H. J. Gierman, W. de Leeuw, O. Giromus, M. H. Indemans, J. Koster, V. Ondrej, R. Versteeg, and R. van Driel (2007). The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. *Molecular and Cellular Biology 27*, 4475–4487.

Green, P. and B. Silverman (1994). *Nonparameteric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman & Hall.

Guelen, L., L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel (2008).

Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature 453*, 948–951.

Hadjur, S. and S. Sofueva (2012). Cohesin-mediated chromatin interactionsinto the third dimension of gene regulation. *Briefings in Functional Genomics 11*, 205–216.

Heard, E. and W. Bickmore (2007). The ins and outs of gene regulation and chromosome territory organisation. *Current Opinion in Cell Biology 19*, 311–316.

Heride, C., M. Ricoul, K. Kiêu, J. von Hase, V. Guillemot, C. Cremer, K. Dubrana, and L. Sabatier (2010). Distance between homologous chromosomes results from chromosome positioning constraints. *Journal of Cell Science 123*, 4063–4075.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Applied Statistics 23*, 340–354.

Hu, M., K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Computational Biology 9*, e1002893.

Imakaev, M., G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods 9*, 999–1003.

Izenman, A. J. (2009). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer.

Kalhor, R., H. Tjong, N. Jayathilaka, F. Alber, and L. Chen (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology 30*, 90–98.

Kato, T. (1966). *Perturbation theory for linear operators*, Volume 132. Springer.

Kent, J., J. Briden, and K. Mardia (1983). Linear and planar structure in ordered multivariate data as applied to progressive demagnetization of palaeomagnetic remanence. *Geophysical Journal of the Royal Astronomical Society 75*, 593–621.

Kruse, K., S. Sewitz, and M. M. Babu (2013). A complex network framework for unbiased statistical analyses of dna–dna contact maps. *Nucleic Acids Research 41*, 701–710.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika 29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika 29*, 115–129.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, J. Dekker, et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science 326*, 289–293.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. California, USA.

Makraz, E. (2010). Apply multidimensional scaling to human chromosome 14. Master's thesis, The University of Leeds, School of Mathematics.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press.

Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods 7*, 1233–1241.

Mateos-Langerak, J., M. Bohn, W. de Leeuw, O. Giromus, E. M. Manders, P. J. Verschure, M. H. Indemans, H. J. Gierman, D. W. Heermann, R. Van Driel, and S. Goetze (2009).

Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences 106*, 3812–3817.

Meluzzi, D. and G. Arya (2013). Recovering ensembles of chromatin conformations from contact probabilities. *Nucleic Acids Research 41*, 63–75.

Oh, M.-S. and A. E. Raftery (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association 96*, 1031–1044.

Peng, C., L.-Y. Fu, P.-F. Dong, Z.-L. Deng, J.-X. Li, X.-T. Wang, and H.-Y. Zhang (2013). The sequencing bias relaxed characteristics of hi-c derived data and implications for chromatin 3d modeling. *Nucleic Acids Research 41*, e183–e183.

Podani, J. and I. Miklos (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology 83*, 3331–3343.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika 42*, 241–266.

Rao, C. (1966). *Linear statistical inference and its applications*. John Wiley.

Rohlf, F. J., T. C. Rodman, and B. J. Flehinger (1980). The use of nonmetric multidimensional scaling for the analysis of chromosomal associations. *Computers and Biomedical Research 13*, 19–35.

Rousseau, M., J. Fraser, M. A. Ferraiuolo, J. Dostie, and M. Blanchette (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using markov chain monte carlo sampling. *BMC Bioinformatics 12*.

Sachs, R., G. Van Den Engh, B. Trask, H. Yokota, and J. Hearst (1995). A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences 92*, 2710–2714.

Sanyal, A., D. Baù, M. A. Martí-Renom, and J. Dekker (2011). Chromatin globules: a common motif of higher order chromosome structure? *Current Opinion in Cell Biology 23*, 325–331.

Segal, M. R., H. Xiong, D. Capurso, M. Vazquez, and J. Arsuaga (2014). Reproducibility of 3d chromatin configuration reconstructions. *Biostatistics*, kxu003.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika 27*, 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika 27*, 219–246.

Shopland, L. S., C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. von Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth, C. Cremer, C. J. Bult, and T. P. O'Brien (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *The Journal of Cell Biology 174*, 27–38.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 217–229.

Sibson, R., A. Bowyer, and C. Osmond (1981). Studies in the robustness of multidimensional scaling: euclidean models and simulation studies. *Journal of Statistical Computation and Simulation 13*, 273–296.

Simonis, M., P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature Genetics 38*, 1348–1354.

Stuart, A. and K. Ord (1994). *Kendall's Advanced Theory of Statistics* (6 ed.), Volume 1. Edward Arnold.

Tanabe, H., S. Müller, M. Neusser, J. von Hase, E. Calcagno, M. Cremer, I. Solovei, C. Cremer, and T. Cremer (2002). Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences 99*, 4424–4429.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika 17*, 401–419.

Trieu, T. and J. Cheng (2014). Large-scale reconstruction of 3d structures of human chromosomes from chromosomal contact data. *Nucleic Acids Research 42*, e52–e52.

Varoquaux, N., F. Ay, W. S. Noble, and J.-P. Vert (2014). A statistical approach for inferring the 3d structure of the genome. *Bioinformatics 30*(12).

Verschure, P. J., I. Van Der Kraan, E. M. Manders, and R. van Driel (1999). Spatial relationship between transcription sites and chromosome territories. *The Journal of Cell Biology 147*, 13–24.

Wächter, A. and L. T. Biegler (2006). On the implementation of an primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming 106*, 25–57.

Yaffe, E. and A. Tanay (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics 43*, 1059–1065.

Young, G. and A. S. Householder (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika 3*, 19–22.