A DYNAMIC STRUCTURE FOR HIGH DIMENSIONAL COVARIANCE MATRICES AND ITS APPLICATION IN PORTFOLIO ALLOCATION

JOHN LEIGH BOX

PhD

UNIVERSITY OF YORK

MATHEMATICS

SEPTEMBER 2015

Abstract

Estimation of high dimensional covariance matrices is an interesting and important research topic. In this thesis, we propose a dynamic structure and develop an estimation procedure for high dimensional covariance matrices. Simulation studies are conducted to demonstrate its performance when the sample size is finite. By exploring a financial application, an empirical study shows that portfolio allocation based on dynamic high dimensional covariance matrices can significantly outperform the market from 1995 to 2014. Our proposed method also outperforms portfolio allocation based on the sample covariance matrix and the portfolio allocation proposed in Fan et al. (2008a).

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to sincerely thank Prof. Wenyang Zhang for his inspiration and guidance throughout this project. He is truly an excellent mentor and I am very grateful for all his time and patience during the last three years.

I would also like to give thanks to all the staff at the Department of Mathematics at the University of York, and in particular my Thesis Advisory Panel members Dr. Marina Knight, Dr. Stephen Connor and Dr. Samer Kharroubi who have given lots of help and encouragement from the start.

I thank my PhD friends Dr Yuan Ke, Mr Xiang Li and Dr Hongjia Yan for countless fascinating discussions and all their advice throughout this project.

Additionally, I would like to acknowledge the York Advanced Research Computing Cluster (YARCC) which I extensively used for the numerical studies in this thesis.

I would also like to acknowledge that my research was supported by an EPSRC funded studentship through the University of York.

Finally I would like to give special thanks to my father Christopher Box, my mother Jenny Box and my girlfriend Zoe-Nicole Manning, for all their love and support.

Author's Declaration

The literature review in Chapter 2 summarises some key ideas related to this thesis. In particular:

- Section 2.1 contains a review of the fundamental concepts in local polynomial modelling and is based on a summary of the first three chapters of the book *Local Polynomial Modelling and its Applications* by Fan and Gijbels (1996).
- Section 2.2 reviews literature concerning varying coefficient models found in *Statistical Estimation in Varying Coefficient Models.* by Fan and Zhang (1999) and *Adaptive Varying-Coefficient Linear Models* by Fan et al. (2003).
- Section 2.3 provides a summary of *High Dimensional Covariance Matrix Estimation using a Factor Model* by Fan et al. (2008a).
- Section 2.4 contains a short summary of modern portfolio theory using ideas originally due to *Portfolio Selection* by Markowitz (1952).

The remaining chapters are related to my submitted paper: *A Dynamic Structure for High Dimensional Covariance Matrices and its Application in Portfolio Allocation* joint with Dr. Shaojun Guo and Prof. Wenyang Zhang.

To the best of my knowledge and belief this thesis does not infringe the copyright of any other person. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

1 Introduction

Covariance matrix estimation is an important topic in statistics and econometrics with wide applications in many disciplines, such as economics, finance and psychology. A traditional approach to estimating covariance matrices is based on the sample covariance matrix. However, the sample covariance matrix would not be a good choice when the dimension is large, and especially when the inverse is required, which is often the case when constructing a portfolio allocation in finance. This is because the estimation errors would accumulate when using the inverse of the sample covariance matrix to estimate the inverse of the covariance matrix. When the size of the covariance matrix is large, the cumulative estimation error would become unacceptable even if the estimation error of each entry of the covariance matrix is tiny.

In recent years there have been various attempts to address high dimensional covariance matrix estimation. Usually, a sparsity condition is imposed to control the trade-off between variance and bias. See, Wu and Pourahmadi (2003), Karoui (2008), Bickel and Levina (2008a,b), Lam and Fan (2009), Fan et al. (2011), and the references therein. Fan et al. (2008a) considered a different approach by imposing a factor model and estimated the covariance matrix based on this structure.

Most of the literature addressing high dimensional covariance matrix estimation assumes that the covariance matrix is constant over time. However, in many applications, covariance matrices are dynamic. For example, today's optimal portfolio allocation may not be optimal tomorrow, or next month. Therefore, when applying the formula for Markowitz's optimal portfolio allocation (Markowitz, 1952, 1968), the covariance matrix used should be dynamic and allowed to change over time.

In order to introduce a dynamic structure for covariance matrices, one cannot simply assume each entry of a covariance matrix is a function of time because this would not serve very well in prediction. Instead, we start with an approach stimulated by Fan et al. (2008a) which is based on the Fama-French three-factor model (Fama and French, 1992, 1993)

$$y_t = \alpha + X_t^{\mathrm{T}}\mathbf{a} + \epsilon_t \tag{1.1}$$

where y_t is the excess return of an asset and X_t is the vector of the three factors at time t. To make (1.1) more flexible, we allow $\mathbf{a}$ to depend on the values of the three factors at time $t-1$. To avoid the so-called 'curse of dimensionality', we assume this dependence is through a linear combination of the values of the three factors at time $t-1$, which brings us to

$$y_t = \alpha(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + X_t^{\mathrm{T}}\mathbf{a}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon_t.$$

This motivates a dynamic structure for the covariance matrix of a random vector Y_t through an adaptive varying coefficient model which we shall now introduce.

Suppose $(X_t^{\mathrm{T}}, Y_t^{\mathrm{T}},\ t = 1, \cdots, n)$ is a time series where Y_t is a p_n dimensional vector and X_t is a q dimensional factor. An underlying assumption throughout this thesis is that $p_n \longrightarrow \infty$ when $n \longrightarrow \infty$, and q is fixed. Also, we assume that $X_t,\ t = 1, \cdots, n$ is a stationary Markov process. We assume

$$Y_t = \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})X_t + \boldsymbol{\epsilon}_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \tag{1.2}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^{\mathrm{T}}$, $\mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})$ is an intercept vector varying with

$X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$, $\boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})$ is a factor loading matrix varying with $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$, and $\{\boldsymbol{\epsilon}_t,\ t = 1, \cdots, n\}$ are random errors which are independent of $\{X_t,\ t = 1, \cdots, n\}$. We also assume that $E(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l : l < t\}) = \mathbf{0}$ and that

$$\mathrm{cov}(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l : l < t\}) = \boldsymbol{\Sigma}_{0,t} = \mathrm{diag}\{\sigma_{1,t}^2, \cdots, \sigma_{p_n,t}^2\}$$

where

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i}\epsilon_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j}\sigma_{k,t-j}^2, \quad t = 2, \cdots, n \tag{1.3}$$

for each $k = 1, \cdots, p_n$ and for some integers m and s. Let $\mathcal{F}_t$ be the σ-algebra generated by $\{(X_l^{\mathrm{T}}, \boldsymbol{\epsilon}_l^{\mathrm{T}}) : \ l \leq t\}$. The main focus of this thesis is on the conditional covariance matrix

$$\mathrm{cov}(Y_t|\mathcal{F}_{t-1}) = \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Sigma}_x(X_{t-1})\boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{\Sigma}_{0,t} \tag{1.4}$$

where $\boldsymbol{\Sigma}_x(X_{t-1}) \equiv \mathrm{cov}(X_t|X_{t-1})$. In (1.4), $\boldsymbol{\beta}$, $\boldsymbol{\Phi}(\cdot)$, $\boldsymbol{\Sigma}_x(\cdot)$, $\alpha_{k,i}$, and $\gamma_{k,j}$ for $i = 0, \cdots, m$ and $j = 1, \cdots, s$ are unknown and need to be estimated. Not only does (1.4) introduce a dynamic structure for $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$, but also reduces the number of unknown parameters from $p_n(p_n+1)/2$ to p_nq+q^2 unknown functions and $q+s+m+1$ unknown parameters.

We remark that model (1.4) is interesting in its own right, since it combines single-index modelling (Carroll et al., 1997, Hardle et al., 1993, Yu and Ruppert, 2002, Xia and Härdle, 2006, Kong et al., 2014) and varying coefficient modelling (Fan and Zhang, 1999, 2000, Fan and Yao, 2003, Sun et al., 2007, Zhang et al., 2009, Li and Zhang, 2011, Sun et al., 2014). In this thesis, as a by-product, an estimation procedure for (1.4) is proposed and an iterative algorithm is developed for implementation purposes.

The organization of this thesis is as follows. In Chapter 2 we review the existing literature related to the proposed methodology such as: local polynomial modelling, varying coefficient models, high dimensional covariance estimation using factor models, and modern portfolio theory. In Chapter 3 we explore a special case of (1.2) corresponding to $p_n = 1$ and two associated methods for estimating $\boldsymbol{\beta}$. This aids as a useful stepping stone to understanding methodology in later chapters, as well as being of independent interest outside the field of covariance matrix estimation. In Chapter 4 we explain how the methodology from Chapter 3 can be generalised when $p_n > 1$. In Chapter 5 we explore the topic of bandwidth selection. In Chapter 6 we propose methodology to estimate $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$ making use of the techniques introduced in Chapters 4 and 5. We provide various simulated data examples to illustrate the performance of the proposed methodology. In Chapter 7 we explore a real data example whereby we use our estimated covariance matrix to form a portfolio and compare its performance with other existing estimators. In Chapter 8 we consider a generalisation to the model structure by modifying the index to take into account $X_{t-1,}$ $X_{t-2}, \cdots, X_{t-\eta}$ for some positive integer η using a moving average. In Chapter 9 we summarise the key conclusions from the thesis and explore a possibility for future work relating to a pursuit of homogeneity.

2 Literature review

2.1 Local polynomial modelling

The main idea behind nonparametric regression is to not assume a parametric form of a regression function. Instead, the functional form of the regression function is left unspecified and determined completely by the data. This approach is useful for getting a clear description of an unknown function, which could suggest whether a parametric choice is appropriate or not. There is clear motivation for the methodology from a least-squares regression point of view, however the whole idea can be extended so that it can be used in: quantile and robust regression, survival analysis, generalized linear models and much more. See, for example, Hastie and Tibshirani (1990), Green and Silverman (1993), Wand and Jones (1994) and Fan and Gijbels (1996) for a comprehensive review of these techniques. In this chapter, we summarise the methodology for local polynomial regression by reviewing Fan and Gijbels (1996).

Assume that we have independently and identically distributed observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ and denote by (X, Y) a generic member of the sample. We start by considering the following global polynomial models

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2.1}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \tag{2.2}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \epsilon_i \tag{2.3}$$

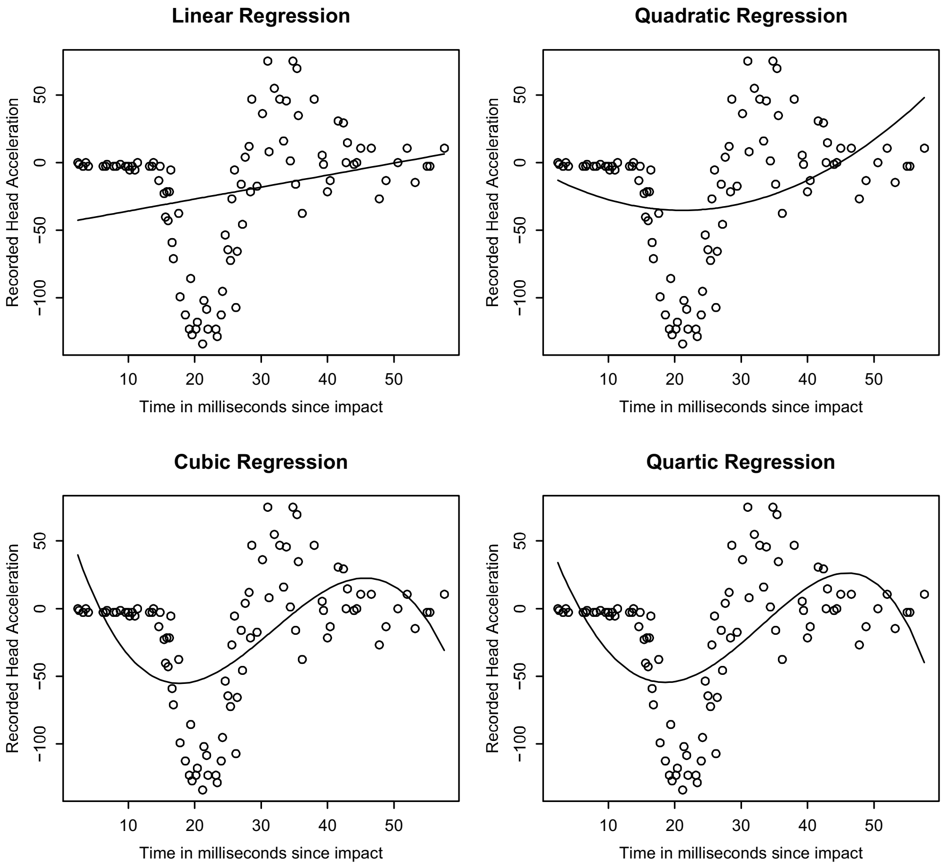
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \epsilon_i \tag{2.4}$$

where $\epsilon_1, \cdots, \epsilon_n$ are independently and identically distributed $N(0, \sigma^2)$ random variables. We define (2.1), (2.2), (2.3) and (2.4) as linear, quadratic, cubic and quartic global polynomial models respectively. Before introducing local polynomial modelling, we shall first consider a motivating example based on the simulated motorcycle accident dataset by Silverman (1985). This dataset contains two variables: *the recorded head acceleration* and *the time in milliseconds since impact,* denoted by Y and X respectively. Figure 1 shows a scatterplot of Y against X along with estimates of $E[Y|X = x]$ resulting from linear, quadratic, cubic and quartic global polynomial models. We can see visually that a linear model is not appropriate due to the lack of linear trend and indeed suffers from large bias. By fitting a quadratic, cubic or quartic regression model, the bias will be reduced but results in an estimator with larger variance. In other words, whichever polynomial order one chooses to parametrize the regression model, the resulting estimator will suffer from large approximation error.

One might argue that it is easy to visualise "using our brains" the relationship between Y and X. This reasoning only holds for datasets within the domain of human visualisation. There are many examples of data where one cannot visualise the data so easily, such as binary data and multivariate data. With that in mind, the purpose of exploring the motorcycle data is simply to give an illustrative example of a problem that local linear regression is particularly well suited for. But applications extend far beyond this simple example and to areas which are beyond the domain of human visualisation.

There are various related methods for fixing the problems resulting from polynomial modelling such as spline approaches and orthogonal series modelling. However, in this chapter and indeed in the entire thesis, we restrict our attention to local polynomial modelling and, in particular, local linear regression. The idea is to apply a (weighted) linear regression model to a strip of data around the point we wish to estimate. We repeat this lots of times over a grid of equally spaced

Figure 1: Example of global polynomial models



Scatterplots of 'recorded head acceleration' against the 'time in milliseconds since impact' from the motorcycle accident dataset along with linear, quadratic, cubic and quartic global polynomial models.

points. Linear interpolation is used for estimation between the grid points. Denote by $m(x) = E(Y|X = x)$ the conditional mean function. We model the data in the strip

$$Y_i = a(x) + b(x)X_i + \text{error}, \qquad \text{for } X_i \in x \pm h \tag{2.5}$$

where h is a prespecified *bandwidth*, and the intercept $a(\cdot)$ and gradient $b(\cdot)$ depend on the grid point x of interest. We choose $a(\cdot)$ and $b(\cdot)$ so that they minimise the weighted local least squares problem

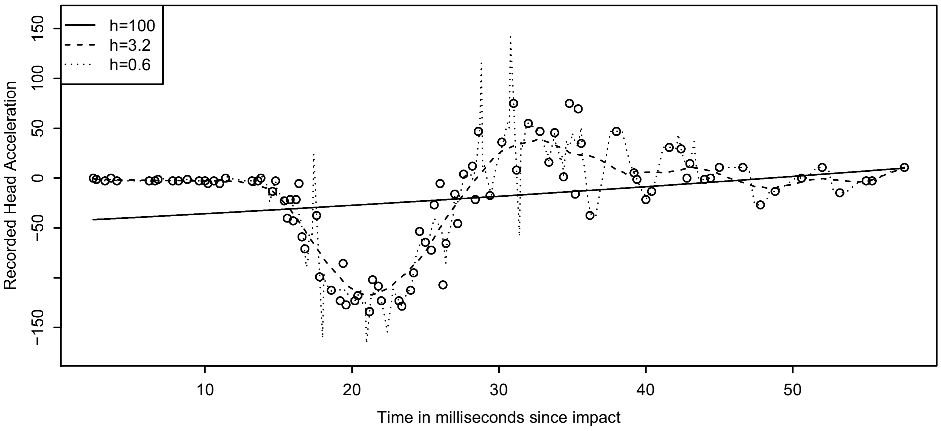
$$\sum_{i=1}^{n}\{Y_i - a(x) - b(x)X_i\}^2 K\left(\frac{X_i - x}{h}\right) I\left(\frac{|X_i - x|}{h} \leq 1\right), \tag{2.6}$$

where $K(\cdot)$ is a *kernel function* (a symmetric probability density function) and I is an indicator function. The kernel function is often chosen to be the Epanechnikov function $K(z) = 0.75 \times (1-z^2)_+$, due to certain desirable statistical properties as explained by J Fan (1995, theorem 3.4). It is for these reasons that we shall always use the Epanechnikov function as the choice of kernel function in this thesis. At this point, it will be useful to define the *scaled kernel function* $K_h(\cdot) = K(\cdot/h)/h$.

One may intuitively understand equation (2.6) as follows. The quantity $K((X_i - x)/h)$ can be thought of as a weight, measuring how far away x is from X_i. That is, $K((X_i - x)/h)$ gets smaller as the distance between x and X_i increases. This is intuitively appealing because remote data points carry little information about our estimate of $m(x)$. One can actually absorb the indicator $I(|X_i - x|/h \leq 1)$ into K if K has a support contained in $[-1, 1]$. The reason we include an indicator is because we wish to discard any data which is outside the strip of data $x \pm h$.

The complexity of the model is determined by the bandwidth h. A very small h results in an estimate which is essentially linear interpo-

Figure 2: Example of local linear modelling



This figure shows a scatter plot of 'recorded head acceleration' against the 'time in milliseconds since impact' from the to the motorcycle accident dataset along with three examples of local linear modelling corresponding to bandwidths $h = 0.6$*,* $h = 3.2$ *and* $h = 100$*.*

lation of the data points. A very large h results in an estimate which coincides with the (global) linear regression estimator. In practice we seek a bandwidth which is just about right and finds a good tradeoff between bias and variance. Sometimes it is possible to test several bandwidths, and choose a bandwidth subjectively (by eye). Another approach is to arbitrarily choose h to be approximately 20% of the range of data. Although this may be sufficient for some purposes, there exist more sophisticated methods for selecting bandwidths such as cross validation or using asymptotic theory.

The motorcycle data can be modelled well using the local linear estimator. Figure 2 shows the local linear model applied to the motorcycle data for a variety of bandwidths. Larger values of h result in higher degrees of smoothing. For example, when $h = 100$ was used, the estimator is almost identical to a global linear model. Conversely, when $h = 0.6$ was used, the estimator essentially interpolates the data points. However, with $h = 3.2$, the local linear fit is just about right, and has a much smaller approximation error than any of the global polynomial fits.

There are various other local modelling regression estimators which one may try and use. For example, one may consider using a weighted average of the response variables

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h\left(X_i - x\right) Y_i}{\sum_{i=1}^{n} K_h\left(X_i - x\right)} \tag{2.7}$$

originally proposed by Nadaraya (1964) and Watson (1964); or the Gasser-Muller estimator

$$\hat{m}_h(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h\left(u - x\right) Y_i \, du \tag{2.8}$$

with $s_i = \left(X_i - X_{i+1}\right)/2$, $X_0 = -\infty$ and $X_{n+1} = +\infty$ originally proposed by Gasser and Müller (1984).

Both (2.7) and (2.8) can be thought of as a local constant approximations for $m(\cdot)$. Indeed, by considering an arbitrary local least squares regression

$$\hat{\theta} = \text{argmin}_{\theta} \sum_{i=1}^{n} (Y_i - \theta)^2 w_i = \sum_{i=1}^{n} w_i Y_i / \sum_{i=1}^{n} w_i \tag{2.9}$$

it is easy to see that (2.7) and (2.8) are special cases with $w_i = K_h\left(X_i - x\right)$ and $w_i = \int_{s_{i-1}}^{s_i} K_h(u - x) du$ respectively. However quite interestingly, local constant fits are rarely used in practice. As explained by Fan (1992), this is because there is no increase in variance when going from a local constant fit to a local linear fit, however there is some reduction in bias. As a matter of fact, there is a stronger generalisation of this very phenomenon, whereby only odd order fits should be used. In addition to this, local linear modelling adapts well to random and fixed designs, as well as highly clustered and nearly uniform designs Fan and Gijbels (1996, chapter 3.2.4).

The local linear estimator can easily be extended to accommodate for a pth order local polynomial fit, as well as νth order derivative estimation. Suppose that the bivariate data $(X_1, Y_1) \cdots (X_n, Y_n)$ form an independent and identically distributed sample from a population (X, Y). We wish to estimate the regression function $m(x_0) = E(Y|X = x_0)$ and its derivatives $m'(x_0), m''(x_0), \cdots, m^{(p)}(x_0)$. Assume that the data is generated from a location-scale model

$$Y = m(X) + \sigma(X)\epsilon \tag{2.10}$$

where $E(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = 1$, and X and ϵ are independent. We denote the conditional variance of Y given $X = x_0$ by $\sigma^2(x_0)$ and the marginal density of X by $f(\cdot)$. Suppose that the $(p+1)$th derivative of $m(x)$ at the point x_0 exists. We start by considering a Taylor expansion of $m(x)$ for x in a neighbourhood of x_0

$$\begin{aligned} m(x) \;&\approx\; m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 \\ &\quad + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \end{aligned} \tag{2.11}$$

From a regression analysis point of view, we can think of $m(x_0)$, $m'(x_0)$, $\cdots$, $m^{(p)}(x_0)$ as unknown model parameters that need to be estimated. This motivates the following notation: let $m^{(j)}(x_0)/j! = \beta_j$ for $j = 0, 1, \cdots, p$. With this in mind, (2.11) becomes

$$m(x) \approx \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_p(x - x_0)^p. \tag{2.12}$$

Comparing (2.12) with (2.5), it is clear that one has the possibility for derivative estimation too. We choose estimators of $\beta_0,\ \beta_1,\ \cdots,\ \beta_p$ so that they minimise

$$\sum_{i=1}^{n}\left\{Y_i-\sum_{j=0}^{p}\beta_j\left(X_i-x_0\right)^j\right\}^2 K_h\left(X_i-x_0\right) \tag{2.13}$$

and denote the estimators by $\hat{\beta}_0,\ \hat{\beta}_1,\ \cdots,\ \hat{\beta}_p$ respectively. Also, we denote the estimator of $m^{(\nu)}(x_0)$ by $\hat{m}_\nu(x_0)=\nu!\hat{\beta}_\nu$ for each $\nu=0,\cdots,p$. The intuition behind equation (2.13) is completely analogous to that of (2.6), just with a higher degree polynomial being fitted locally rather than a straight line. Traditionally with least squares theory, it is convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix of problem (2.13)

$$\mathbf{X}=\begin{pmatrix} 1 & (X_1-x_0) & \cdots & (X_1-x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n-x_0) & \cdots & (X_n-x_0)^p \end{pmatrix},$$

and define

$$\mathbf{y}=(Y_1,\cdots,Y_n)^{\mathrm{T}},\quad \boldsymbol{\beta}=(\beta_0,\cdots,\beta_p)^{\mathrm{T}},\quad \mathbf{W}=\operatorname{diag}\left\{K_h(X_i-x_0)\right\}.$$

The weighted least squares problem (2.13) can be rewritten as

$$\min_{\boldsymbol{\beta}}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{W}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)$$

for which the estimator has the following analytic formula:

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}. \tag{2.14}$$

Since both the bias and variance of an estimator determine the mean squared error (MSE) and the mean integrated squared error (MISE), it is natural to be interested in their properties. From a practical point of view, one way of selecting a bandwidth is so that MISE is as small as possible. The conditional expectation and variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{m} \\ &= \boldsymbol{\beta} + (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{r} \end{aligned}$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = \left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^{\mathrm{T}}\Sigma\mathbf{X}\right)\left(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X}\right)^{-1}$$

where $\mathbf{m} = \{m(X_1), \cdots, m(X_n)\}^{\mathrm{T}}$, $\boldsymbol{\beta} = \left\{m(x_0), \cdots, m^{(p)}(x_0)/p!\right\}^{\mathrm{T}}$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\boldsymbol{\beta}$, the vector of residuals of the local polynomial approximation, and $\Sigma = \mathrm{diag}\left\{K_h^2\left(X_i - x_0\right)\sigma^2(X_i)\right\}$. However, notice that these equations cannot be directly used because of the unknown quantities $\mathbf{r}$ and Σ. Therefore a first order asymptotic expansion of the bias and variance of $\hat{m}_\nu(x_0) = \nu!\hat{\boldsymbol{\beta}}_\nu$ is used as an approximation and is given in the theorem below. The theorem is quoted directly from Fan and Gijbels (1996) but was originally proven by Ruppert and Wand (1994). We use the following notation:

$$\mu_j = \int u^j K(u)du, \quad \nu_j = \int u^j K^2(u)du, \quad S = (u_{j+l})_{0\leq j,l\leq p},$$

$$\tilde{S} = (u_{j+l+1})_{0\leq j,l\leq p}, \quad S^* = (u_{j+l})_{0\leq j,l\leq p}, \quad c_p = \left(\mu_{p+1}, \cdots, \mu_{2p+1}\right)^{\mathrm{T}},$$

$$\tilde{c}_p = (\mu_{p+2}, \cdots, \mu_{2p+2})^{\mathrm{T}}, \quad e_{\nu+1} = (0, \cdots, 0, 1, 0, \cdots, 0)^{\mathrm{T}},$$

where $e_{\nu+1}$ has a 1 on the $(\nu+1)^{th}$ position. Also we use $o_P(1)$ to represent a random quantity that is tending to zero in probability.

Theorem

Assume that $f(x_0) > 0$ and that $f(\cdot)$, $m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighbourhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\hat{m}_\nu(x_0)$ is given by

$$\begin{aligned} Var(\hat{m}_\nu(x_0)|\mathbb{X}) &= e_{\nu+1}^{\mathrm{T}} S^{-1} S^* S^{-1} e_{v+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} \\ &\quad + o_P\left(\frac{1}{nh^{1+2\nu}}\right). \end{aligned} \tag{2.15}$$

The asymptotic conditional bias for $p - \nu$ odd is given by

$$\begin{aligned} Bias\{\hat{m}_\nu(x_0)|\mathbb{X}\} &= e_{\nu+1}^{\mathrm{T}} S^{-1} c_p \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-\nu} \\ &\quad + o_P(h^{p+1-\nu}). \end{aligned} \tag{2.16}$$

Further, for $p - \nu$ even the asymptotic conditional bias is

$$\begin{aligned} Bias\{\hat{m}_\nu(x_0)|\mathbb{X}\} &= e_{\nu+1}^{\mathrm{T}} S^{-1} \tilde{c}_p \frac{\nu!}{(p+1)!} \{m^{(p+2)}(x_0), \\ &\quad + (p+2) m^{(p+1)}(x_0) \frac{f'(x_0)}{f(x_0)} \} h^{p+2-\nu} \\ &\quad + o_P(h^{p+2-\nu}) \end{aligned} \tag{2.17}$$

provided that $f'(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighbourhood of x_0 and $nh^3 \to \infty$.

From the above theorem, we can clearly see that there is a theoretical distinction between odd order fits and even order fits with respect to the asymptotic bias. Indeed, Ruppert and Wand (1994) showed that in fact odd order fits are always desirable over even order fits.

2.2 Varying coefficient models

A varying coefficient model (VCM) is, as the name suggests, an extension of a linear model where the coefficients are allowed to vary over some random variable U. Their modelling potential has been explored, for example, by Hastie and Tibshirani (1990), Cleveland et al. (1992) and Fan and Zhang (1999). In this section, we shall provide a concise review of VCM's.

We assume the following conditional linear structure

$$Y = \sum_{j=1}^{p} a_j(U) X_j + \epsilon \tag{2.18}$$

for given covariates $(U, X_1, \cdots, X_p)^{\mathrm{T}}$ and response variable Y with

$$E(\epsilon | U, X_1, \cdots, X_p) = 0, \quad Var(\epsilon | U, X_1, \cdots, X_p) = \sigma^2(U).$$

A model of this structure is known as a *varying coefficient model.* Note that it is possible for us to include an intercept by setting, for example, the first variable $X_1 \equiv 1$. Also, note that the coefficient functions $a_j(\cdot)$ vary over a known random variable U. One major advantage of this, is that it helps to reduce modelling bias by avoiding the curse of dimensionality. The coefficient functions a_j often have a nice interpretability, especially in longitudinal data analysis where they represent how the impact of the corresponding covariate on the response changes over time. Suppose we have a random sample $(U_i, X_{i1}, \cdots, X_{ip}, Y_i)$ where $i = 1, \cdots, n$ from model (2.18), then one can estimate the coefficient functions $a_j(\cdot)$ $j = 1, \cdots, p$ using local linear modelling. As explained by Fan and Zhang (1999), for each given u we approximate the function locally by

$$a_j(U_i) \approx a_j + b_j(U_i - u)$$

for U_i in a neighbourhood of u. This leads to the following loss function

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p} \{a_j + b_j(U_i - u)\} X_{ij} \right]^2 K_h(U_i - u). \tag{2.19}$$

The weighted least squares problem (2.19) can be rewritten as

$$\min_{\mathbf{a}} \left(\mathcal{Y} - \mathcal{X}\mathbf{a}\right)^{\mathrm{T}} \mathcal{W} \left(\mathcal{Y} - \mathcal{X}\mathbf{a}\right)$$

where $\mathbf{a} = (a_1, b_1, \cdots, a_p, b_p)^{\mathrm{T}}$ and

$$\mathcal{X} = \begin{pmatrix} X_{11} & X_{11}(U_1 - u) & \cdots & X_{1p} & X_{1p}(U_1 - u) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ X_{n1} & X_{n1}(U_n - u) & \cdots & X_{np} & X_{np}(U_n - u) \end{pmatrix}, \quad \mathcal{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\mathcal{W} = \operatorname{diag}\left\{K_h(U_1 - u), \cdots, K_h(U_n - u)\right\}.$$

The solution is given by

$$\mathbf{a} = (\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{Y}$$

and the estimator of coefficient function $a_j(u)$ is

$$\hat{a}_j(u) = e_{2j-1,2p}^{\mathrm{T}}(\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathrm{T}}\mathcal{W}\mathcal{Y} \tag{2.20}$$

where $e_{k,m}$ is the unit vector of length m with the k-th component being 1.

In traditional varying coefficient models the index U is known and often chosen to be a time component. However in the model structure studied in this thesis, the index is unknown and the estimation procedure becomes more complex. Suppose we wish to estimate a multivariate regression function $G(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ where Y is a random variable and $\mathbf{X}$ is a $p \times 1$ random vector. The following model structure, termed an *adaptive varying-coefficient linear model* (AVCLM) in Fan et al. (2003), is one way to approximate $G(\mathbf{x})$

$$g(\mathbf{x}) = \sum_{j=0}^{p} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})x_j \tag{2.21}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown direction, $\mathbf{x} = (x_1 \cdots x_p)^{\mathrm{T}}$, $x_0 = 1$ and coefficients $g_0(\cdot), \cdots .g_p(\cdot)$ are unknown functions. The choice for the estimators of $g_j(\cdot)$ and $\boldsymbol{\beta}$ are based on the minimisation of $E\{G(\mathbf{X}) - g(\mathbf{X})\}^2$. One example of an estimation procedure for $\boldsymbol{\beta}$ was given in Fan et al. (2003), and a thorough discussion on this topic is provided in Section 3.2.1. As a by-product of this thesis, we shall propose an improved estimation procedure for AVCLM's. It is useful to note that once $\boldsymbol{\beta}$ has successfully been estimated, model (2.21) becomes a VCM which can be estimated using local linear regression.

2.3 High Dimensional Covariance matrix estimation

We shall review the topic of high dimensional covariance matrix estimation, that is when the number of dimensions p is comparable to sample size n. This has applications in finance but also in other fields such as economics and psychology. In the context of modern portfolio theory, suppose that $(\mathbf{Y}_1, \cdots, \mathbf{Y}_n)$ is a time series, where $\mathbf{Y}_t = (Y_{t1}, \cdots, Y_{tp})^{\mathrm{T}}$ represents the excess return of an asset over the risk-free rate of return.

One way to estimate the covariance matrix $\mathbf{\Sigma}$ is by using the sample covariance matrix. It has been shown that the sample covariance matrix is unbiased and invertible when $p < n$, and is a good choice when there is no prior knowledge of the underlying covariance structure (Eaton and Tyler, 1994). However, when p grows with (or even exceeds) n, it has been shown by Ledoit and Wolf (2004) and Johnstone (2001) that the sample covariance matrix no longer performs well. This is due to noise accumulation, whereby small element-wise estimation errors accumulate so that the matrix as a whole becomes a poor estimator. One way to deal with this problem is to impose a factor structure on the response variable Y_i. This is an interesting possibility if one has prior understanding of the underlying structure of the data. This was explored by Fan et al. (2008a) and we shall now give a concise review of the topic.

Suppose that

$$Y_i = b_{i1}f_1 + \cdots + b_{iK}f_K + \epsilon_i, \qquad i = 1, \cdots, p, \tag{2.22}$$

where $f_1, \cdots, f_K$ are the excess returns of K factors, b_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, K$, are unknown factor loadings, and $\epsilon_1, \cdots, \epsilon_p$ are p idiosyncratic errors uncorrelated given $f_1, \cdots, f_K$. We assume that the factors are observable. A real life example of this comes from the Fama and French Three Factor Model Fama and French (1993), whereby $K = 3$ and the factors $f_1,\ f_2,\ f_3$ denote proxies for market, size and value factors. Indeed, it is well known that the variability in stock returns can be explained well by these three factors. Hence, aided with this prior knowledge, it is possible to estimate $\mathbf{\Sigma}$ as follows. First, we write the factor model (2.22) in matrix form

$$\mathbf{y} = \mathbf{B}_n\mathbf{f} + \boldsymbol{\epsilon}, \tag{2.23}$$

where

$$\mathbf{y} = (Y_1, \cdots, Y_p)^{\mathrm{T}}, \quad \mathbf{B}_n = (\mathbf{b}_1, \cdots, \mathbf{b}_p)^{\mathrm{T}},$$

$$\mathbf{b}_i = (b_{n,i1}, \cdots, b_{n,iK})^{\mathrm{T}}, \quad \mathbf{f} = (f_1, \cdots, f_K)^{\mathrm{T}}, \quad \boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_p)^{\mathrm{T}}$$

for $i = 1, \cdots, p$. This notation is chosen to be consistent with that of Fan et al. (2008a). We remark that the factor loading matrix $\mathbf{B}_n$ has a subscript n to emphasise the dependence on n. It is assumed that

$$E(\boldsymbol{\epsilon}|\mathbf{f}) = \mathbf{0}, \quad \text{and,} \quad \mathrm{cov}(\boldsymbol{\epsilon}|\mathbf{f}) = \boldsymbol{\Sigma}_{n,0}$$

where $\boldsymbol{\Sigma}_{n,0}$ a diagonal matrix. Let $(\mathbf{f}_1, \mathbf{y}_1), \cdots, (\mathbf{f}_n, \mathbf{y}_n)$ be n independent and identically distributed samples of $(\mathbf{f}, \mathbf{y})$, and let

$$\boldsymbol{\Sigma}_n = \mathrm{cov}(\mathbf{y}),\ \mathbf{X} = (\mathbf{f}_1, \cdots, \mathbf{f}_n),\ \mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n) \text{ and } \mathbf{E} = (\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n).$$

Under model (2.23) we have

$$\boldsymbol{\Sigma}_n = \mathrm{cov}(\mathbf{B}_n\mathbf{f}) + \mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{B}_n\mathrm{cov}(\mathbf{f})\mathbf{B}_n^{\mathrm{T}} + \boldsymbol{\Sigma}_{n,0}.$$

One way to estimate the covariance matrix $\boldsymbol{\Sigma}_n$ is with the substitution estimator

$$\hat{\boldsymbol{\Sigma}}_n = \hat{\mathbf{B}}_n\widehat{\mathrm{cov}}(\mathbf{f})\hat{\mathbf{B}}_n^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}_{n,0}$$

where

$$\hat{\mathbf{B}}_n = \mathbf{Y}\mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}})^{-1}$$

is the matrix of estimated regression coefficients;

$$\widehat{\mathrm{cov}}(\mathbf{f}) = (n-1)^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}} - \{n(n-1)\}^{-1}\mathbf{X}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}$$

is the sample covariance matrix of the factors $\mathbf{f}$; and

$$\hat{\mathbf{\Sigma}}_{n,0} = \mathrm{diag}(n^{-1}\hat{\mathbf{E}}\hat{\mathbf{E}}^{\mathrm{T}})$$

is the diagonal matrix of $n^{-1}\hat{\mathbf{E}}\hat{\mathbf{E}}^{\mathrm{T}}$ with $\hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{B}}\mathbf{X}$ the matrix of residuals.

The covariance matrix estimator proposed in this thesis is significantly different to $\hat{\mathbf{\Sigma}}_n$ because the structure we propose is dynamic and we do not assume the factors are constant. By assuming that the covariance matrix is dynamic, we will later see that significant improvements can be made when applied to real data.

2.4 Modern portfolio theory

Some of the first quantitative research on portfolio allocation was originally analysed by Markowitz (1952), which we shall now review. Suppose that $\mathbf{w} = (w_1, \cdots, w_p)^{\mathrm{T}}$ represents weights of a portfolio corresponding to p assets. We assume that the weights can be negative, which means short selling is allowed. Further we assume that the weights must sum to one, so they have the interpretation of an allocation. Markowitz defines the mean-variance optimal portfolio as the solution to the following minimization problem

$$\begin{aligned} &\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{w} \\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1} = 1 \quad \text{and} \quad \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \delta \end{aligned} \tag{2.24}$$

where $\mathbf{\Sigma}$ is the covariance matrix of excess asset returns and $\boldsymbol{\mu}$ is the expected vector of asset returns, $\mathbf{1}$ is a $p \times 1$ vector of ones, and δ is the target rate of return imposed on the portfolio. This has the interpretation that we wish to minimise the portfolio's variance subject to a desired return. It is easy to see that an analytic solution of (2.24) is given by

$$\mathbf{w} = \frac{c_3 - c_2\gamma}{c_1c_3 - c_2^2}\boldsymbol{\Sigma}^{-1}\mathbf{1} + \frac{c_1\gamma - c_2}{c_1c_3 - c_2^2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{2.25}$$

with: $c_1 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{1}$, $c_2 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, and $c_3 = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

There are other portfolio allocations which can be constructed by solving different optimisation problems, such as a minimum variance portfolio

$$\begin{aligned}&\min_{\mathbf{w}} \ \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}\\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1} = 1,\end{aligned} \tag{2.26}$$

a mean-variance optimal portfolio with no short sales

$$\begin{aligned}&\min_{\mathbf{w}} \ \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}\\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1} = 1, \quad \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \delta, \quad \text{and } w_i \geq 0 \text{ for all } i,\end{aligned} \tag{2.27}$$

and a mean-variance portfolio with gross exposure constraints

$$\begin{aligned}&\min_{\mathbf{w}} \ \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}\\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1}_{p_n} = 1, \ \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \delta \text{ and } \ \|\mathbf{w}\|_1 \leq c\end{aligned} \tag{2.28}$$

where c is some constant controlling the overall exposure of the portfolio.

The crucial point is that in order to form a portfolio allocation $\mathbf{w}$, using any of the above approaches, one needs to obtain an estimate of $\boldsymbol{\Sigma}$ which is difficult when the dimension p is large relative to n. Furthermore, today's optimal portfolio allocation may not be optimal tomorrow. Hence, in this thesis we shall estimate both $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ using a proposed dynamic structure and make comparisons with the traditional approach based on the sample covariance matrix and the factor model explored by Fan et al. (2008a). In the real data analysis of this thesis, we will calculate the returns resulting from the portfolio allocations (2.24) and (2.28) to assess the performance of the estimators.

3 Univariate model

In this chapter we explore a special case of model (1.2), which can be understood as a univariate model corresponding to the case when $p_n = 1$. The purpose of exploring this univariate model is twofold. First, it helps us to gain insight into estimation procedures which will be later generalised in Chapter 4. Secondly, it is of independent interest outside the field of covariance matrix estimation, since the proposed methodology can be used for AVCLM's in general. For example, AVCLM's have previously been used to analyse Canadian mink–muskrat data in 1925–1994 and the pound–dollar exchange rates in 1974–1983 by Fan et al. (2003). Using a simulation study, the method we propose will be shown to have a better performance.

3.1 Model specification

Assume that $\{(X_t, y_t),\ t = 1, \cdots, n\}$ is a time series where y_t denotes a univariate response variable and $X_t = (x_{1,t}, \cdots, x_{q,t})^{\mathrm{T}}$ is a random vector. We assume that $\{X_t,\ t = 1, \cdots, n\}$ is a stationary Markov process, and consider a model of the form

$$y_t = \sum_{j=0}^{q} g_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t} + \epsilon_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \tag{3.1}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^{\mathrm{T}}$ is an unknown direction vector; $g_0(\cdot)$,$\cdots$,$g_q(\cdot)$ are unknown coefficient functions varying with scalar index $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$; $x_{0,t} = 1$ is an intercept dummy variable; and ϵ_t is a random error variable. In this chapter, we simplify the GARCH structure (1.3) and assume that $\{\epsilon_t,\ t = 1, \cdots, n\}$ are independent random variables with mean zero and variance σ^2.

We remark that in a similar way to an AVCLM, we must impose the identifiability conditions that $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. Note, however, that (3.1) is similar, but not identical, to an AVCLM. Recall that in an AVCLM the impact of $x_{j,t}$ on y_t depends on $g_j(X_t^{\mathrm{T}}\boldsymbol{\beta})$ as opposed $g_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})$. This subtle change in the model structure means we no longer have to impose the identifiability condition $g_q(\cdot) \equiv 0$. See the proof of theorem 1(b) in Fan et al. (2003) for an explanation why.

3.2 Methodology

In this section we outline two methods for estimating $\boldsymbol{\beta}$ in model (3.1). One crucial thing to note is that once an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given, model (3.1) becomes a varying coefficient model with known index $X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. At this stage, estimation of coefficient functions $g_j(\cdot)$ can be achieved using methodology associated with traditional varying coefficient models using local linear regression. This can be achieved by approximating $g_j(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, $j = 0, \cdots, q$, locally by a Taylor expansion

$$g_j(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \approx g_j(z) + \dot{g}_j(z)(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)$$

for $X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ in a neighbourhood of a given grid point z. By minimising

$$\sum_{t=2}^{n}(y_t - \sum_{j=0}^{q}\{g_j(z) + \dot{g}_j(z)(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)\}x_{j,t})^2 K_h(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)$$

with respect to $g_j(z)$ and $\dot{g}_j(z)$ for $j = 0, \cdots, q$, it follows from least squares theory that

$$\hat{\boldsymbol{\theta}} = \{\mathcal{X}^{\mathrm{T}}W\mathcal{X}\}^{-1}\mathcal{X}^{\mathrm{T}}W\mathbf{y} \tag{3.2}$$

where

$$\hat{\boldsymbol{\theta}} = (\hat{g}_0(z), \cdots, \hat{g}_q(z), \hat{\dot{g}}_0(z), \cdots, \hat{\dot{g}}_q(z))^{\mathrm{T}}, \tag{3.3}$$

$$W = \mathrm{diag}\left(K_h(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z), \cdots, K_h(X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)\right), \tag{3.4}$$

$$\mathcal{X} = \begin{pmatrix} 1 & X_2^{\mathrm{T}} & X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z & (X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)X_2^{\mathrm{T}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{\mathrm{T}} & X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z & (X_{n-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)X_n^{\mathrm{T}} \end{pmatrix}, \tag{3.5}$$

and $\mathbf{y} = (y_2, \cdots, y_n)^{\mathrm{T}}$.

With this in mind, the primary focus now is the estimation of $\boldsymbol{\beta}$. Due to the complex relationship between Y_t and X_t, there is not an analytic formula for an estimator of $\boldsymbol{\beta}$, and hence we explore two iterative approaches in Section 3.2.1 and Section 3.2.2.

3.2.1 Fan's estimator for $\boldsymbol{\beta}$

We shall now introduce an estimator $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ of $\boldsymbol{\beta}$ analogous to that originally suggested by Fan et al. (2003).

Before the iterative procedure, one must specify an initial value of $\boldsymbol{\beta}$, which we denote by $\tilde{\boldsymbol{\beta}}$. The problem of choosing $\tilde{\boldsymbol{\beta}}$ was not discussed in Fan et al. (2003), but it was said that "[they] expect that the estimator derived will be good if the initial value is reasonably good." However, we will see in the simulation study that an incorrect choice of $\tilde{\boldsymbol{\beta}}$ can slightly worsen the performance of the estimator.

Define the residual sum of squares of the data by

$$R(\boldsymbol{\beta}) = \frac{1}{n-1}\sum_{t=2}^{n}\{y_t - \sum_{j=0}^{q} g_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\}^2 w(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) \tag{3.6}$$

where $w(\cdot)$ is a bounded weight function with a bounded support. In practice, we usually let $w(\cdot)$ be an indicator function in order to reduce the boundary effects. In the simulation studies, the choice of $w(\cdot)$ is not crucial because the results are not highly sensitive to its particular choice. Hence, in the numerical examples, we set

$$w(z) = \begin{cases} 1 & \text{if} \quad \kappa_{2.5}(X_1^{\mathrm{T}}\boldsymbol{\beta}, \cdots, X_{n-1}^{\mathrm{T}}\boldsymbol{\beta}) \le z \le \kappa_{97.5}(X_1^{\mathrm{T}}\boldsymbol{\beta}, \cdots, X_{n-1}^{\mathrm{T}}\boldsymbol{\beta}) \\ 0 & \text{otherwise} \end{cases}$$

where $\kappa_p(X_1^{\mathrm{T}}\boldsymbol{\beta}, \cdots, X_{n-1}^{\mathrm{T}}\boldsymbol{\beta})$ is the pth percentile of $X_1^{\mathrm{T}}\boldsymbol{\beta}, \cdots, X_{n-1}^{\mathrm{T}}\boldsymbol{\beta}$. The quantity $R(\boldsymbol{\beta})$ can be interpreted as a goodness of fit statistic, and can be used to estimate $\boldsymbol{\beta}$ by iterating between the following two steps until $R(\boldsymbol{\beta})$ differs insignificantly.

Step 1: Estimate $g_j(\cdot)$ assuming $\boldsymbol{\beta}$ is known In this step we assume $\boldsymbol{\beta}$ is known to us and estimate $g_m(\cdot)$. For example if this is the first iteration set $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$, otherwise set $\hat{\boldsymbol{\beta}}$ equal to the estimator of $\boldsymbol{\beta}$ obtained from Step 2 of the previous iteration. Since an estimate for $\boldsymbol{\beta}$ has been given, model (3.1) becomes a synthetic varying coefficient model with known index $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$. Consequently, estimates of the coefficient functions and their derivatives can be obtained from local linear regression using (3.2).

Step 2: Estimate $\boldsymbol{\beta}$ assuming $g_j(\cdot)$ is known Using $\hat{\boldsymbol{\theta}}$, the estimated coefficient functions and their derivatives obtained from Step 1, the goal now is to estimate $\boldsymbol{\beta}$. Suppose that $\boldsymbol{\beta}_{(1)}$ is the true minimizer of (3.6). Unfortunately it is not possible to find an exact analytic formula for $\boldsymbol{\beta}_{(1)}$. However, by noting that $\dot{\mathbf{R}}(\boldsymbol{\beta}_{(1)}) = 0$, where $\dot{\mathbf{R}}(\cdot)$ denotes the derivative of $R(\cdot)$, we can obtain an approximate minimizer. For any $\boldsymbol{\beta}_{(0)}$ close to $\boldsymbol{\beta}_{(1)}$, we have the Taylor expansion

$$\mathbf{0} = \dot{R}(\boldsymbol{\beta}_{(1)}) \approx \dot{R}(\boldsymbol{\beta}_{(0)}) + \ddot{R}(\boldsymbol{\beta}_{(0)})(\boldsymbol{\beta}_{(1)} - \boldsymbol{\beta}_{(0)}),$$

where $\ddot{\mathbf{R}}(\cdot)$ is the Hessian matrix of $R(\cdot)$. Upon re-arrangement, this suggests a one-step iterative estimator

$$\hat{\boldsymbol{\beta}}_{(1)} = \boldsymbol{\beta}_{(0)} - \ddot{R}(\boldsymbol{\beta}_{(0)})^{-1}\dot{R}(\boldsymbol{\beta}_{(0)}),$$

where $\hat{\boldsymbol{\beta}}_{(1)}$ is the new estimate for $\boldsymbol{\beta}$. Then, rescale $\hat{\boldsymbol{\beta}}_{(1)}$ such that it has unit norm with first non-vanishing element positive. Note that we have the following expressions for the gradient vector and Hessian:

$$\begin{aligned}\dot{R}(\boldsymbol{\beta}) &= -\frac{2}{n-1}\sum_{t=2}^{n}\{Y_t - \sum_{j=0}^{q} g_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\}\{\sum_{j=0}^{q}\dot{g}_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\} \\ &\quad \times X_{t-1}w(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\end{aligned}$$

and

$$\begin{aligned}\ddot{R}(\boldsymbol{\beta}) &= \frac{2}{n-1}\sum_{t=2}^{n}\{\sum_{j=0}^{q}\dot{g}_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\}^2 X_{t-1}X_{t-1}^{\mathrm{T}}w(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) \\ &\quad - \frac{2}{n-1}\sum_{t=2}^{n}\{Y_t - \sum_{j=0}^{q} g_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\} \\ &\quad \times \{\sum_{j=0}^{q}\ddot{g}_j(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})x_{j,t}\}X_{t-1}X_{t-1}^{\mathrm{T}}w(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})\end{aligned}$$

where we assume the derivative of the weight function $w(\cdot)$ is zero for simplicity.

3.2.2 Local discrepancy estimator

In this section, we propose an alternative estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ of $\boldsymbol{\beta}$ which we will later show outperforms $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ when applied to simulated data.

For $m = 0, \cdots, q$, expanding $g_m(\cdot)$ locally using a first order Taylor series gives us the following approximation:

$$g_m(X_i^{\mathrm{T}}\boldsymbol{\beta}) \approx g_m(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \dot{g}_m(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_i^{\mathrm{T}}\boldsymbol{\beta} - X_j^{\mathrm{T}}\boldsymbol{\beta})$$

where $X_i^{\mathrm{T}}\boldsymbol{\beta}$ is in a small neighbourhood of $X_j^{\mathrm{T}}\boldsymbol{\beta}$. Therefore we can approximate model (3.1) by

$$y_i \approx \sum_{m=0}^{q} g_m(X_j^{\mathrm{T}}\boldsymbol{\beta})x_{m,i} + \sum_{m=0}^{q} \dot{g}_m(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_{i-1}^{\mathrm{T}}\boldsymbol{\beta} - X_j^{\mathrm{T}}\boldsymbol{\beta})x_{m,i} + \epsilon_t. \tag{3.7}$$

For brevity purposes, we use the notation

$$\begin{aligned}\boldsymbol{\gamma}_j &= (g_j^{(0)}, \cdots, g_j^{(q)}, \dot{g}_j^{(0)}, \cdots, \dot{g}_j^{(q)})^{\mathrm{T}} \\ \boldsymbol{\Gamma} &= (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_{n-1}).\end{aligned}$$

where we denote $g_m(X_j^{\mathrm{T}}\boldsymbol{\beta})$ and $\dot{g}_m(X_j^{\mathrm{T}}\boldsymbol{\beta})$ by $g_j^{(m)}$ and $\dot{g}_j^{(m)}$ respectively. Using approximation (3.7) together with the idea of least squares, we can form the following loss function:

$$\begin{aligned}L(\boldsymbol{\beta}, \boldsymbol{\Gamma}) \ &= \ \sum_{j=1}^{n-1}\sum_{i=2}^{n}\{y_i - (\sum_{m=0}^{q} g_j^{(m)}x_{m,i} + \sum_{m=0}^{q} \dot{g}_j^{(m)}x_{m,i}(X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta})\}^2 \\ &\quad \times K_h((X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta}) \qquad (3.8)\end{aligned}$$

which we call the *local discrepancy loss function.* The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ can be obtained by solving

$$(\hat{\boldsymbol{\beta}}_{\mathrm{L}}, \hat{\boldsymbol{\Gamma}}) = \underset{\boldsymbol{\beta},\boldsymbol{\Gamma}}{\operatorname{argmin}} \ L(\boldsymbol{\beta}, \boldsymbol{\Gamma}) \tag{3.9}$$

subject to the constraints $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. A global minimum of the local discrepancy function cannot be found analytically and therefore an iterative procedure is proposed for implementation purposes. It is worth noting that at each stage of the iterative procedure, there exists a closed form solution. This means the estimation procedure does not depend on a generic maximisation or minimisation algorithm.

As with Fan's methodology, introduced in Section 3.2.1, one must choose an initial estimator for $\boldsymbol{\beta}$ which we still denote by $\tilde{\boldsymbol{\beta}}$. Not only will we later see that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ outperforms $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ when applied to simulated data, but also the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is not very sensitive to the initial estimator $\tilde{\boldsymbol{\beta}}$.

In order to solve equation (3.9), we iterate between the following two steps until $L(\boldsymbol{\beta}, \boldsymbol{\Gamma})$ differs insignificantly.

Step 1: Estimate $\boldsymbol{\Gamma}$ assuming $\boldsymbol{\beta}$ is known If this is the first iteration, set $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. Otherwise, set $\hat{\boldsymbol{\beta}}$ equal to the estimator of $\boldsymbol{\beta}$ obtained from Step 2 of the previous iteration. We now estimate $\boldsymbol{\Gamma}$ by solving

$$\hat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma}}{\operatorname{argmin}}\{L(\hat{\boldsymbol{\beta}}, \boldsymbol{\Gamma})\}.$$

For each j, we choose the estimator of $\boldsymbol{\gamma}_j$ by minimising

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j}{\operatorname{argmin}} \sum_{i=2}^{n} \{y_i - \sum_{m=0}^{q} g_j^{(m)} x_{m,i} - \sum_{m=0}^{q} \dot{g}_j^{(m)} x_{m,i} (X_{i-1} - X_j)^{\mathrm{T}} \hat{\boldsymbol{\beta}}\}^2 \\ \times K_h((X_{i-1} - X_j)^{\mathrm{T}} \hat{\boldsymbol{\beta}}). \qquad (3.10)\end{aligned}$$

We can write equation (3.10) in matrix notation as

$$\hat{\boldsymbol{\gamma}}_j = \underset{\boldsymbol{\gamma}_j}{\operatorname{argmin}} \; (\mathbf{y} - \mathcal{X}_j \boldsymbol{\gamma}_j)^{\mathrm{T}} W_j (\mathbf{y} - \mathcal{X}_j \boldsymbol{\gamma}_j)$$

where

$$W_j = \operatorname{diag}\{K_h((X_1 - X_j)^{\mathrm{T}} \hat{\boldsymbol{\beta}}), \; \cdots, \; K_h((X_{n-1} - X_j)^{\mathrm{T}} \hat{\boldsymbol{\beta}})\}, \qquad (3.11)$$

$$\mathcal{X}_j = \begin{pmatrix} 1 & X_2^{\mathrm{T}} & (X_1 - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}} & X_2^{\mathrm{T}}(X_1 - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n^{\mathrm{T}} & (X_{n-1} - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}} & X_n^{\mathrm{T}}(X_{n-1} - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}} \end{pmatrix}, \quad (3.12)$$

and $\mathbf{y} = (y_2, \cdots, y_n)^{\mathrm{T}}$. By least squares theory, the solution is given by

$$\hat{\boldsymbol{\gamma}}_j = \{\mathcal{X}_j^{\mathrm{T}} W_j \mathcal{X}_j\}^{-1} \mathcal{X}_j^{\mathrm{T}} W_j \mathbf{y}$$

and hence we obtain $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \cdots, \hat{\boldsymbol{\gamma}}_{n-1})$.

Step 2: Estimate $\boldsymbol{\beta}$ assuming $\boldsymbol{\Gamma}$ is known Using the estimates $\hat{\boldsymbol{\Gamma}}$, from Step 1, we would like to choose our estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{(1)}$, such that

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{L(\boldsymbol{\beta}, \hat{\boldsymbol{\Gamma}})\}$$

which is equivalent to

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{(1)} \; = \; & \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{ \sum_{j=1}^{n-1}\sum_{i=2}^{n}\{y_i - \sum_{m=0}^{q} \hat{g}_j^{(m)} x_{m,i} \\ & - \sum_{m=0}^{q} \hat{g}_j^{(m)} x_{m,i}(X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta}\}^2 K_h((X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta})\Big\}. \quad (3.13) \end{aligned}$$

However, notice that $\boldsymbol{\beta}$ appears twice in this objective function: once in the least squares part and once in the kernel function. It is unlikely that a closed form exists for equation (3.13), and so we use the following approximation

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(1)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{j=1}^{n-1}\sum_{i=2}^{n}\{y_i - \sum_{m=0}^{q}\hat{g}_j^{(m)}x_{m,i} \\ &\quad - \sum_{m=0}^{q}\hat{g}_j^{(m)}x_{m,i}(X_{i-1}-X_j)^{\mathrm{T}}\boldsymbol{\beta}\}^2 K_h((X_{i-1}-X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(0)})\}.\end{aligned} \tag{3.14}$$

where $\hat{\boldsymbol{\beta}}_{(0)}$ is the estimator for $\boldsymbol{\beta}$ used in Step 1. We remark that $\boldsymbol{\beta}$ only appears once in this objective function, in the least squares part, and so a closed form solution can be obtained as follows. First, rewrite the minimisation problem:

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{j=1}^{n-1}\sum_{i=2}^{n}\left\{c_{ij} - M_{ij}^{\mathrm{T}}\boldsymbol{\beta}\right\}^2 w_{i-1,j} \tag{3.15}$$

where:

$$c_{ij} = y_i - \sum_{m=0}^{q}\hat{g}_j^{(m)}x_{m,i}, \quad M_{ij} = \sum_{m=0}^{q}\hat{g}_j^{(m)}x_{m,i}(X_{i-1}-X_j),$$

$$w_{ij} = K_h((X_i - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(0)}).$$

All that remains is to rewrite the double summation in equation (3.15) in the form of a single summation of $N = (n-1)(n-1)$ quantities, so that it becomes a traditional weighted least squares problem. We do this by stacking the vectors and matrices. This can be achieved by defining the $N \times 1$ vector $\mathbf{C}$, the $N \times q$ matrix $\mathbf{M}$, the $N \times N$ diagonal matrix $\mathbf{W}$ as follows:

$$\begin{aligned}
\mathbf{C} &= \left(c_{21}, \cdots, c_{n1}, c_{22}, \cdots, c_{n2}, \cdots, c_{2,(n-1)}, \cdots, c_{n,(n-1)}\right)^{\mathrm{T}} \\
\mathbf{M} &= \left(M_{21}, \cdots, M_{n1}, M_{22}, \cdots, M_{n2}, \cdots, M_{2,(n-1)}, \cdots, M_{n,(n-1)}\right)^{\mathrm{T}} \\
\mathbf{W} &= \operatorname{diag}\{w_{21}, \cdots, w_{n1}, w_{22}, \cdots, w_{n2}, \cdots, w_{2,n-1}, \cdots, w_{n,n-1}\}.
\end{aligned}$$

With this notation, equation (3.15) can be written as a traditional weighted least squares problem:

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \{\mathbf{C}_{[i]} - \mathbf{M}_{[i]}\boldsymbol{\beta}\}^2 \mathbf{W}_{[i]}$$

where $\mathbf{M}_{[i]}$ and $\mathbf{C}_{[i]}$, denote the ith row of $\mathbf{M}$ and $\mathbf{C}$ respectively and where $\mathbf{W}_{[i]}$ denotes the ith diagonal entry. In matrix notation, this is equivalent to

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathbf{C} - \mathbf{M}\boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{C} - \mathbf{M}\boldsymbol{\beta}\right),$$

and the solution is given by

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{C}. \tag{3.16}$$

At this point, $\hat{\boldsymbol{\beta}}_{(1)}$ should be rescaled to satisfy the identifiability conditions $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. Although we now have an analytic formula for $\hat{\boldsymbol{\beta}}_{(1)}$, one should proceed with caution to avoid exceeding RAM limitations because the dimensions of $\mathbf{C}$, $\mathbf{M}$ and $\mathbf{W}$ increase at rate $O(n^2)$. This issue becomes more crucial in the multivariate analogue, and a recommended solution is given in Section 4.2.1.

3.3 Simulation study

In this section, we are going to use a simulated example to compare the performance of estimators $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$. We generate 1000 datasets from model (3.1) each with a sample size of $n = 1000$. We set

$$q = 4, \quad \boldsymbol{\beta} = \frac{1}{3}(1,\ 2,\ 0,\ 2)^{\mathrm{T}}, \quad g_0(z) = 3\mathrm{exp}(-z^2),$$

$$g_1(z) = 0.8z, \quad g_2(z) = 0, \quad g_3(z) = 1.5\mathrm{sin}(\pi z), \quad g_4(z) = 0.$$

In other words, we generate y_t, $t = 1, \cdots, n$, from the following model

$$y_t = 3\mathrm{exp}(-z_{t-1}^2) + 0.8z_{t-1}x_{1,t} + 1.5\mathrm{sin}(\pi z_{t-1})x_{3,t} + \epsilon_t$$

where

$$z_t = \frac{1}{3}(x_{1,t} + 2x_{2,t} + 2x_{4,t})$$

and where we generate $X_0, \cdots, X_n$ independently from a uniform distribution on $[-1, 1]^q$ and $\epsilon_1, \cdots, \epsilon_n$ independently from a standard normal distribution. The parameter $\boldsymbol{\beta}$ and the true coefficient functions are chosen to be analogous to Example 1 in Fan et al. (2003).

For each generated dataset, we estimate $\boldsymbol{\beta}$ using $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$. In each case, we try the following two initial estimators

$$\tilde{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}, \quad \text{and} \quad \tilde{\boldsymbol{\beta}}_2 = \frac{1}{\sqrt{166}}(8,\ 1,\ -10,\ 1)^{\mathrm{T}}. \tag{3.17}$$

We remark that $\tilde{\boldsymbol{\beta}}_1$ is a perfect choice since it conveniently coincides with the true value of $\boldsymbol{\beta}$ which aids the estimation procedure. However, the angle between $\tilde{\boldsymbol{\beta}}_2$ and the true $\boldsymbol{\beta}$ is approximately 36 degrees, and therefore not an optimal choice. We remark that in other simulated work, not reported in this thesis for brevity, we tried other initial estimators which do not point close to the true $\boldsymbol{\beta}$. We find that the same conclusions hold, and hence $\tilde{\boldsymbol{\beta}}_2$ can be thought of as a representative example.

In addition to this we estimate $\boldsymbol{\beta}$ using the "Oracle" estimator $\hat{\boldsymbol{\beta}}_{\mathrm{O}} = \boldsymbol{\beta}$ which uses the true $\boldsymbol{\beta}$ as the estimator. Hence for each estimate of $\boldsymbol{\beta}$, ($\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, using $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$) we estimate coefficient functions $g_j(\cdot)$, $j = 0, \cdots, q$, over 101 equally spaced grid points of the following range

$$r(\hat{\boldsymbol{\beta}}) = \max(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \min(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}). \tag{3.18}$$

We denote these grid points by $u_1, \cdots, u_{101}$.

In the iterative algorithms for estimating $\boldsymbol{\beta}$, we choose the bandwidth h equal to 20% of $r(\tilde{\boldsymbol{\beta}})$ on the first iteration and update it on subsequent iterations by choosing h equal to 20% of $r(\hat{\boldsymbol{\beta}}_{(0)})$ where $\hat{\boldsymbol{\beta}}_{(0)}$ is the most recent estimate. After the iterative procedure converges, and we have an estimate of $\boldsymbol{\beta}$, we estimate $g_j(\cdot)$ using h equal to 20% of $r(\hat{\boldsymbol{\beta}})$. We opt for this approach to simplify this simulation study. However, a significant improvement, along with a thorough topic of bandwidth selection, can be found in Chapter 5.

Evaluating the performance of $\hat{\boldsymbol{\beta}}$ and $\hat{g}_j$, $j = 0, \cdots, q$ is done using the relative absolute deviation error metrics

$$\Delta(\hat{g}) = \frac{\sum_{j=0}^{q}\sum_{k=1}^{101} |\hat{g}_j(u_k) - g_j(u_k)|}{\sum_{j=0}^{q}\sum_{k=1}^{101} |g_j(u_k)|} \qquad \Delta(\hat{\boldsymbol{\beta}}) = \frac{\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|_1}{\left\|\boldsymbol{\beta}\right\|_1} \tag{3.19}$$

where $\left\|\cdot\right\|_1$ denotes the L_1 norm. The expectation and standard deviation of $\Delta(\hat{g})$ and $\Delta(\hat{\boldsymbol{\beta}})$ can be approximated by averaging over the 1000 replications using

$$E(\Delta(\hat{g})) \approx \frac{1}{1000}\sum_{i=1}^{1000} \Delta_i(\hat{g}), \;\; \mathrm{SD}(\Delta(\hat{g})) \approx \big(\frac{1}{1000}\sum_{i=1}^{1000}\{\Delta_i(\hat{g}) - E(\Delta(\hat{g}))\}^2\big)^{1/2}$$

and

$$E(\Delta(\hat{\boldsymbol{\beta}})) \approx \frac{1}{1000}\sum_{i=1}^{1000}\Delta_i(\hat{\boldsymbol{\beta}}),\ \ \mathrm{SD}(\Delta(\hat{\boldsymbol{\beta}})) \approx \big(\frac{1}{1000}\sum_{i=1}^{1000}\{\Delta_i(\hat{\boldsymbol{\beta}}) - E(\Delta(\hat{\boldsymbol{\beta}}))\}^2\big)^{1/2}$$

where $\Delta_i(\cdot)$ denotes the relative error metric of the ith simulated dataset.

A comparison of the expectation and standard deviation of $\Delta(\hat{\boldsymbol{\beta}})$ and $\Delta(\hat{g})$ over the 1000 replications is given in Table 3.1, and we see that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ performs significantly better than $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$. Further we see that $E(\Delta(\hat{g}))$ and $\mathrm{SD}(\Delta(\hat{g}))$ achieved by $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is very close to that achieved by $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$.

As one would expect, choosing the initial value $\tilde{\boldsymbol{\beta}}$ so that it points closer to the true value can also lead to a reduction in approximation error for both estimators. This is best illustrated by the boxplots in Figure 3. For the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, a key observation is that only two datasets out of 1000 suffered significantly by choosing $\tilde{\boldsymbol{\beta}}_2$ instead of $\tilde{\boldsymbol{\beta}}_1$. However, for the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, the median relative error increased from approximately 30% to 40% when moving from $\tilde{\boldsymbol{\beta}}_1$ to $\tilde{\boldsymbol{\beta}}_2$. This shows that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is notably less sensitive to the initial estimator than $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ is. This is very important in real data applications because in reality we do not know the true $\boldsymbol{\beta}$ and we cannot guarantee that our initial estimator is well chosen.

The boxplots in Figure 3 also give us insight into the reason why $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ fails to perform as well as $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$. We see the distributions of $\Delta(\hat{\boldsymbol{\beta}})$ and $\Delta(\hat{g})$ are positively skewed for $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, with larger errors usually corresponding to the iterative algorithm failing to converge or converging to local optima. However, for $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, the iterative algorithm converged to the global optimum 998 out of 1000 times.

Indeed, if the estimation of $\boldsymbol{\beta}$ is poor, then so is the estimation of the coefficient functions, due to the nature of the model structure. This is illustrated in Figures 5, 6, 7, 8 and 9 which show typical estimated coefficient functions resulting from $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, using $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ respectively. A vital observation is that the typical estimated coefficient functions, corresponding to the seed matching the upper quartile of $\Delta(\hat{g})$, can be seen to have significantly large approximation error when $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ is used. This is because the estimator failed to converge or converged to a local optimum. However for $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ the typical estimated coefficient functions always seem to closely approximate the true function. It is also worth noting that the estimated coefficient functions resulting from $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ look fairly similar to those from $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$. This shows that the performance of the proposed estimator is almost as good as the true $\boldsymbol{\beta}$ from a visual point of view. Although there are sometimes deviations at the boundary, this is a fairly common issue in nonparametric regression due to the lack of information available at the boundary.

It is also important to make a comparison of the computational time to estimate $\boldsymbol{\beta}$ because the methodology in future chapters will be based on a generalisation of this one. From Table 3.2 we see that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is significantly faster on average than $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$. In both cases, the computational time decreases if $\tilde{\boldsymbol{\beta}}$ is chosen to be $\tilde{\boldsymbol{\beta}}_1$ instead of $\tilde{\boldsymbol{\beta}}_2$, as one would naturally expect. This is because fewer iterations are needed for convergence.

To conclude, we have firstly shown that the proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ outperforms the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ suggested by Fan et al. (2003) in this simulation study. Secondly, we have seen evidence to suggest that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is not very sensitive to the choice of the initial estimator $\tilde{\boldsymbol{\beta}}$. Finally, we see that the performance of $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is similar to that of the Oracle estimator $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$.

Table 3.1: Comparison of estimators of $\boldsymbol{\beta}$

	$E(\Delta(\hat{\boldsymbol{\beta}}))$	$\mathrm{SD}(\Delta(\hat{\boldsymbol{\beta}}))$	$E(\Delta(\hat{g}))$	$\mathrm{SD}(\Delta(\hat{g}))$
$\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_1$	0.926	1.038	0.482	0.299
$\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_2$	1.069	1.062	0.523	0.309
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$	0.063	0.028	0.262	0.043
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$	0.067	0.098	0.264	0.055
$\hat{\boldsymbol{\beta}}_{\mathrm{O}}$	0.000	0.000	0.261	0.042

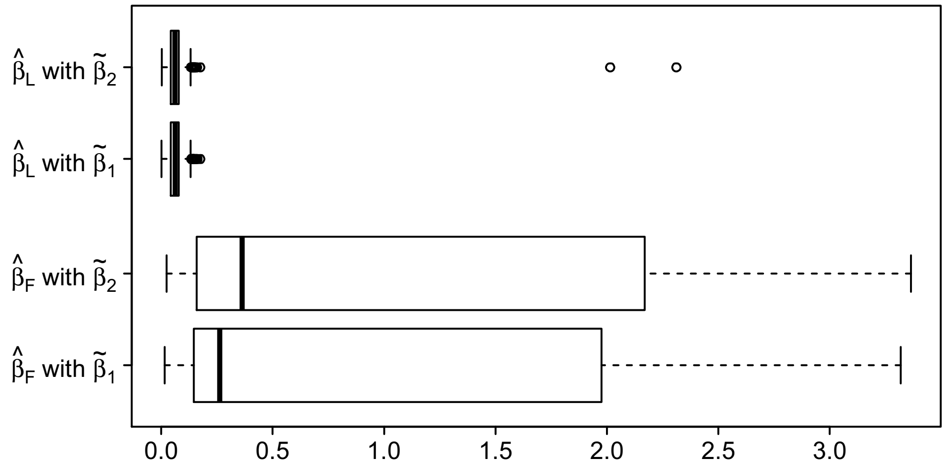
This table gives a comparison of the estimators $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ in terms of $E(\Delta(\hat{\boldsymbol{\beta}}))$, $\mathrm{SD}(\Delta(\hat{\boldsymbol{\beta}}))$, $E(\Delta(\hat{g}))$, and $\mathrm{SD}(\Delta(\hat{g}))$ for the simulation study in Section 3.3. For both $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, we compare the performance when the initial values $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ are used at the beginning of the iterative estimation procedures.

Table 3.2: Computational time to estimate $\boldsymbol{\beta}$

	Mean	Standard Deviation
$\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_1$	9.82	2.05
$\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_2$	10.62	2.67
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$	4.31	1.28
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$	7.56	2.07

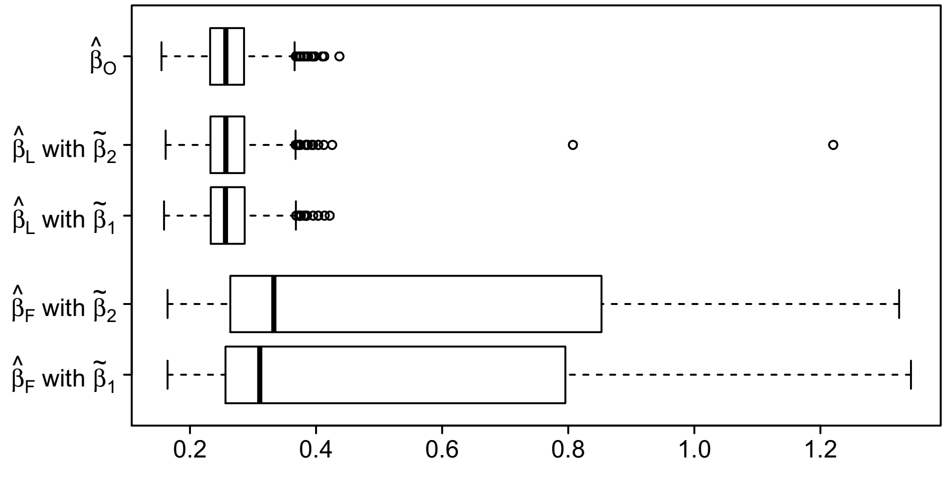
For the simulation study in Section 3.3, this table shows the mean and standard deviation of the computational time (seconds) spent estimating $\boldsymbol{\beta}$ for a single dataset. Results are given for estimators $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ with initial values $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$.

Figure 3: Boxplots of $\Delta_i(\hat{\boldsymbol{\beta}})$



This figure shows the boxplots of the relative error metrics $\Delta_i(\hat{\boldsymbol{\beta}})$, $i = 1, \cdots, 1000$, *for* $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ *using* $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, *using* $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ *using* $\tilde{\boldsymbol{\beta}}_1$, *and* $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ *using* $\tilde{\boldsymbol{\beta}}_2$ *for the simulation study in Section 3.3.*

Figure 4: Boxplots of $\Delta_i(\hat{g})$



This figure shows the boxplots of the relative error metrics $\Delta_i(\hat{g})$, $i = 1, \cdots, 1000$, *for* $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ *using* $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, *using* $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ *using* $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ *using* $\tilde{\boldsymbol{\beta}}_2$, *and* $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$ *for the simulation study in Section 3.3.*

Figure 5: Estimated coefficients with $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_1$.

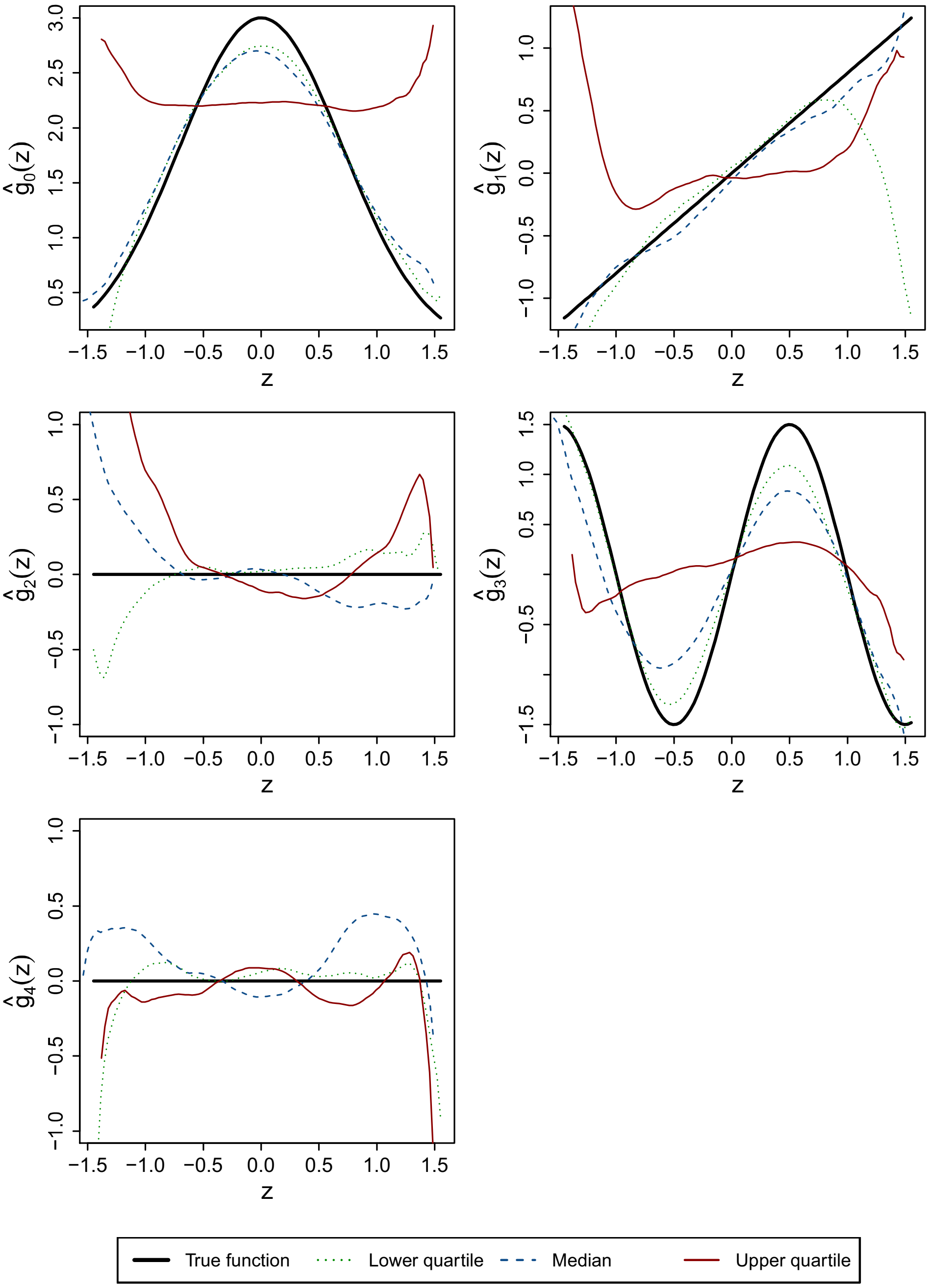


Figure 6: Estimated coefficients with $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$.

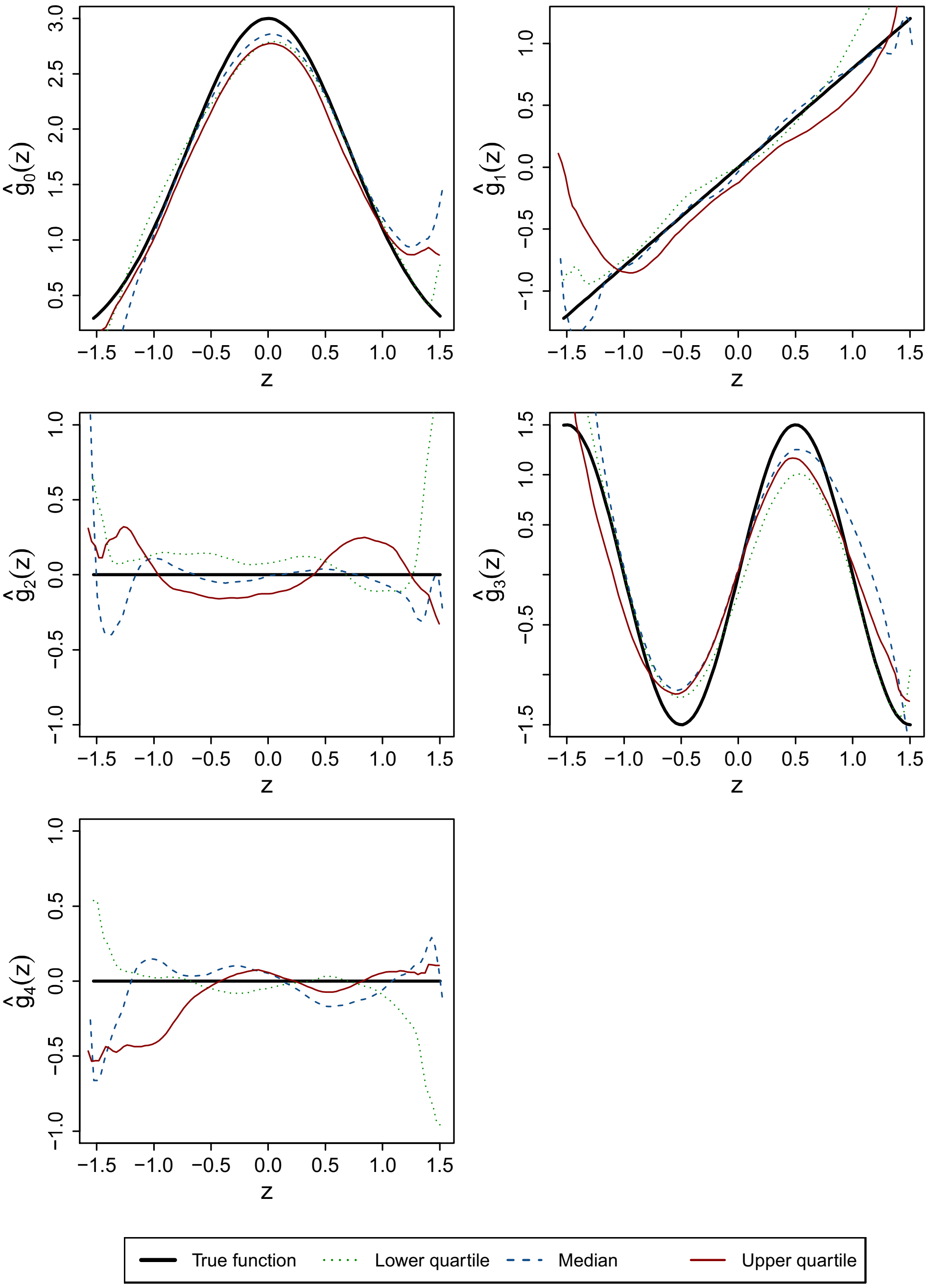


Figure 7: Estimated coefficients with $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$ using $\tilde{\boldsymbol{\beta}}_2$

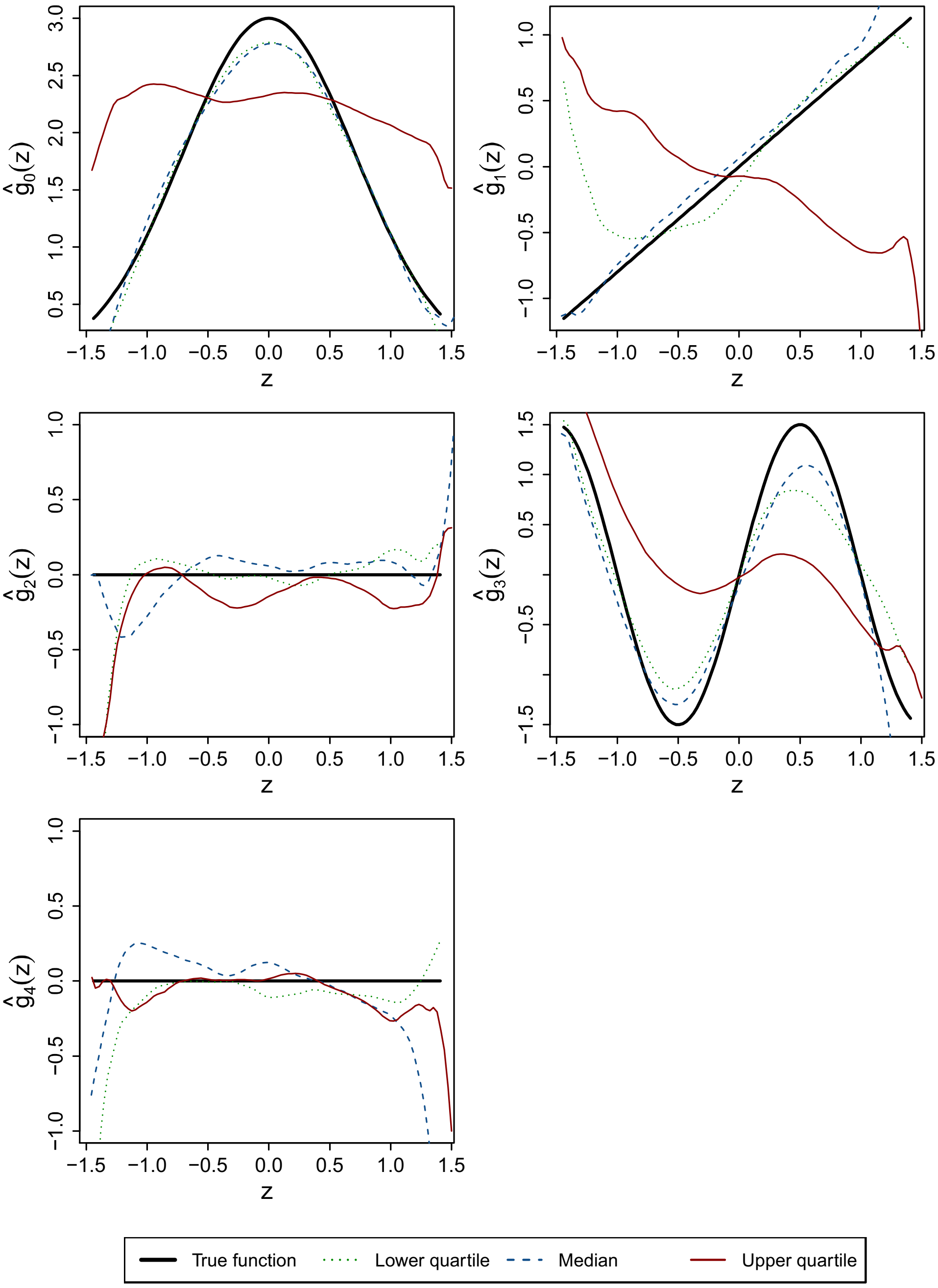


Figure 8: Estimated coefficients with $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$.

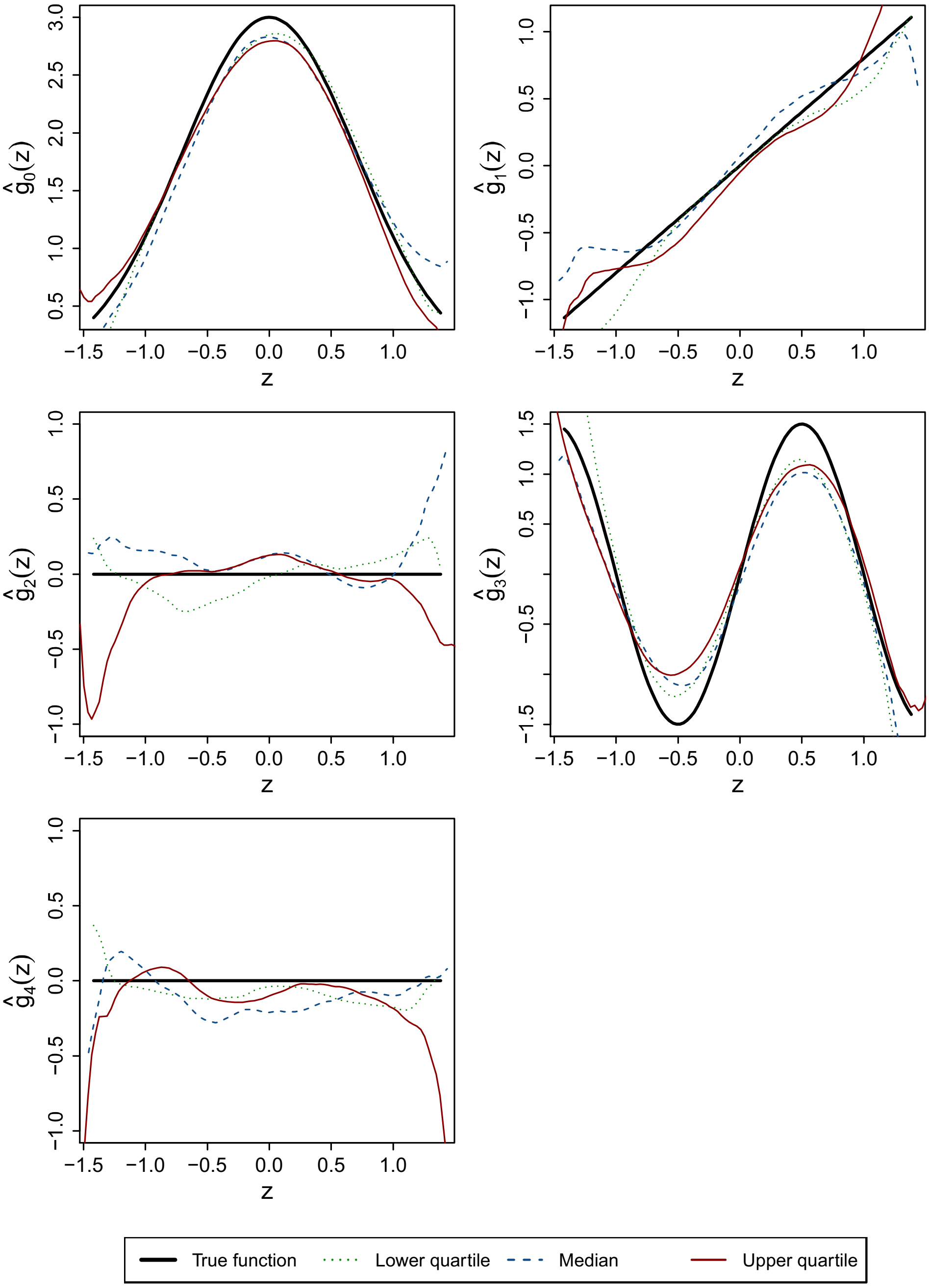
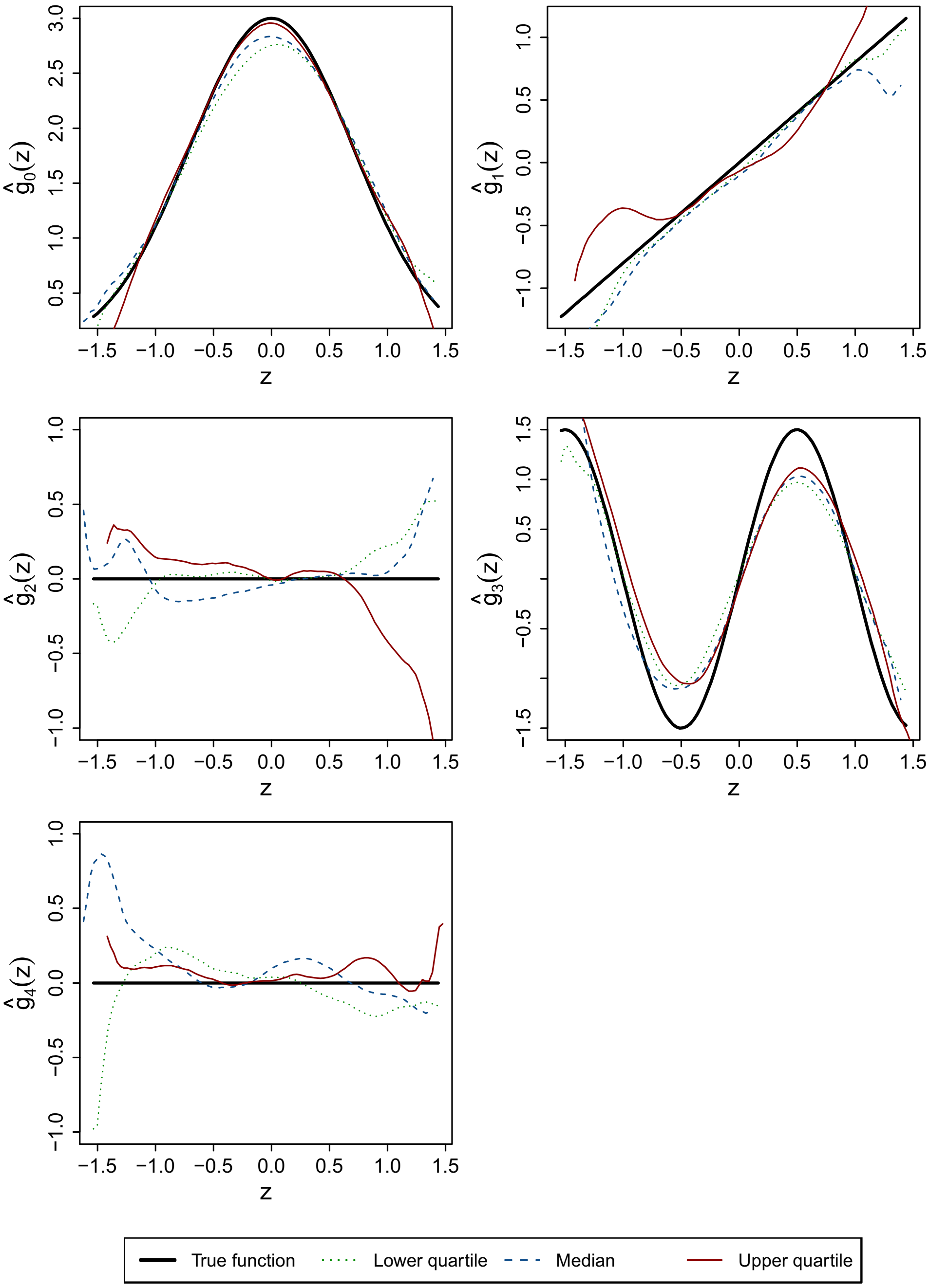


Figure 9: Estimated coefficients with $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$.



4 Multivariate model

In this chapter we introduce a multivariate generalisation of model (3.1). The main objective is to explore three methods for estimating $\boldsymbol{\beta}$ in this multivariate model, and use a simulation study to compare the performance of these estimators.

4.1 Model specification

Assume that $\{(X_t, Y_t),\ t = 1, \cdots, n\}$ is a time series where Y_t denotes a vector of p_n response variables and X_t denotes a vector of q (observable) factors. We assume that $p_n \longrightarrow \infty$ as $n \longrightarrow \infty$, and q is fixed. We also assume that $\{X_t,\ t = 1, \cdots, n\}$ is a stationary Markov process, and consider the following model structure

$$Y_t = \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})X_t + \boldsymbol{\epsilon}_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \qquad (4.1)$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^{\mathrm{T}}$ is an unknown direction vector, $\mathbf{g}(\cdot)$ is an unknown intercept vector, $\boldsymbol{\Phi}(\cdot)$ is an unknown factor loading matrix, and $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \cdots, \epsilon_{p_n,t})^{\mathrm{T}}$ is a p_n-dimensional random error vector at time t. In this chapter, we assume that $\{\epsilon_{k,t},\ t = 1, \cdots, n\}$ are independent variables with mean zero and variance σ^2 for each k.

We can write $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ as

$$\mathbf{g}(\cdot) = (g_1(\cdot), \cdots, g_{p_n}(\cdot))^{\mathrm{T}}, \quad \text{and} \quad \boldsymbol{\Phi} = (\mathbf{a}_1(\cdot), \cdots, \mathbf{a}_{p_n}(\cdot))^{\mathrm{T}}$$

where $g_j(\cdot)$ and $\mathbf{a}_j^{\mathrm{T}}(\cdot)$ are defined as the rows of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ respectively. This allows us to rewrite (4.1) using componentwise notation

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon_{k,t}, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \quad (4.2)$$

where

$$Y_t = (y_{1,t}, \cdots, y_{p_n,t})^{\mathrm{T}}, \quad X_t = (x_{1,t}, \cdots, x_{q,t})^{\mathrm{T}}.$$

It is easy to see that (3.1) is a special case of (4.2) when $p_n = 1$. An important remark, however, is that $\boldsymbol{\beta}$ does not depend on k. In other words, the components of the model all share the same $\boldsymbol{\beta}$. The reason for making this assumption becomes clear if one were to consider a slightly different model

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + \epsilon_{k,t}, \tag{4.3}$$

where $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,1})^{\mathrm{T}}$ and $\|\boldsymbol{\beta}_k\| = 1$, $\beta_{k,1} > 0$ for $k = 1, \cdots, p_n$. The key problem with (4.3) is that we have an extra $(p_n - 1)q$ parameters to estimate, which could result in a significant increase in variance in the estimation. It is for this reason that we assume (4.2) in this chapter and the rest of the thesis. An alternative possibility for future work, however, is discussed in Section 9.2.

4.2 Methodology

In a similar way to Section 3.2, once an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given, model (4.1) becomes equivalent to p_n synthetic varying coefficient models with known index $X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. For any z, the estimators of $g_k(z)$ and $\mathbf{a}_k(z)$ can be obtained using local linear estimation for standard varying-coefficient models

$$\hat{g}_k(z) = (1, \mathbf{0}_{1\times(2q+1)})\{\mathcal{X}^{\mathrm{T}}W\mathcal{X}\}^{-1}\mathcal{X}^{\mathrm{T}}W\mathbf{y}_k \tag{4.4}$$

and

$$\hat{\mathbf{a}}_k(z) = (\mathbf{0}_{q\times 1}, I_q, \mathbf{0}_{q\times(q+1)})\{\mathcal{X}^{\mathrm{T}}W\mathcal{X}\}^{-1}\mathcal{X}^{\mathrm{T}}W\mathbf{y}_k \tag{4.5}$$

where $\mathbf{y}_k = (y_{k,2}, \cdots, y_{k,n})^{\mathrm{T}}$, and where W and $\mathcal{X}$ are defined in equation (3.4) and equation (3.5) respectively. However, the methodology for estimating $\boldsymbol{\beta}$ is notably different from Section 3.2. Hence, the objective of this section is to introduce three methods for estimating $\boldsymbol{\beta}$.

4.2.1 Local discrepancy estimator

In this section we propose an estimator for $\boldsymbol{\beta}$ which is based on a multivariate generalisation of the estimator given in Section 3.2.2.

A Taylor expansion, for $X_i^{\mathrm{T}}\boldsymbol{\beta}$ in a small neighbourhood of $X_j^{\mathrm{T}}\boldsymbol{\beta}$, gives

$$a_{k,m}(X_i^{\mathrm{T}}\boldsymbol{\beta}) \approx a_{k,m}(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \dot{a}_{k,m}(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta}$$

and

$$g_k(X_i^{\mathrm{T}}\boldsymbol{\beta}) \approx g_k(X_j^{\mathrm{T}}\boldsymbol{\beta}) + \dot{g}_k(X_j^{\mathrm{T}}\boldsymbol{\beta})(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta}$$

where $\mathbf{a}_k^{\mathrm{T}}(\cdot) = (a_{k,1}(\cdot), \cdots, a_{k,q}(\cdot))$, for $m = 1, \cdots, q$ and $k = 1, \cdots, p_n$. We rewrite this as:

$$g_{k,i}^{(m)} \approx g_{k,j}^{(m)} + \dot{g}_{k,j}^{(m)}(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta}$$

where

$$g_{k,j}^{(0)} = g_k(X_j^{\mathrm{T}}\boldsymbol{\beta}),\ \dot{g}_{k,j}^{(0)} = \dot{g}_k(X_j^{\mathrm{T}}\boldsymbol{\beta}),\ g_{k,j}^{(m)} = a_{k,m}(X_j^{\mathrm{T}}\boldsymbol{\beta}),\ \dot{g}_{k,j}^{(m)} = \dot{a}_{k,m}(X_j^{\mathrm{T}}\boldsymbol{\beta}).$$

We can approximate model (4.2) by

$$y_{k,i} \approx \sum_{m=0}^{q} g_{k,j}^{(m)} x_{m,i} + \sum_{m=0}^{q} \dot{g}_{k,j}^{(m)} x_{m,i}(X_i - X_j)^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{k,i}$$

where $x_{0,i} = 1$ for all i. To keep future equations concise, we use the notation

$$\boldsymbol{\gamma}_{k,j} = (g_{k,j}^{(0)}, \cdots, g_{k,j}^{(q)}, \dot{g}_{k,j}^{(0)}, \cdots, \dot{g}_{k,j}^{(q)})^{\mathrm{T}}$$

$$\mathbf{\Gamma}_k = \{\boldsymbol{\gamma}_{k,1}, \cdots, \boldsymbol{\gamma}_{k,n-1}\}, \quad \mathbf{\Gamma} = \{\mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_{p_n}\}.$$

This leads to the multivariate local discrepancy loss function

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{\Gamma}) \; = \; & \sum_{j=1}^{n-1} \sum_{i=2}^{n} \sum_{k=1}^{p_n} \{y_{k,i} - \sum_{m=0}^{q} g_{k,j}^{(m)} x_{m,i} \\ & - \sum_{m=0}^{q} \dot{g}_{k,j}^{(m)} x_{m,i} (X_{i-1} - X_j)^{\mathrm{T}} \boldsymbol{\beta}\}^2 K_h((X_{i-1} - X_j)^{\mathrm{T}} \boldsymbol{\beta}). \end{aligned} \tag{4.6}$$

As with the univariate case, the estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, can be obtained by solving

$$(\hat{\boldsymbol{\beta}}_{\mathrm{L}}, \hat{\mathbf{\Gamma}}) = \underset{\boldsymbol{\beta}, \mathbf{\Gamma}}{\operatorname{argmin}} \; L(\boldsymbol{\beta}, \mathbf{\Gamma}) \tag{4.7}$$

subject to the constraints $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$. The estimation procedure given in Section 3.2.2 is a special case of what is to follow. Choose an initial estimator for $\boldsymbol{\beta}$ which we denote by $\tilde{\boldsymbol{\beta}}$ and iterate between the following two steps until $L(\boldsymbol{\beta}, \mathbf{\Gamma})$ differs insignificantly.

Step 1: Estimate $\mathbf{\Gamma}$ assuming $\boldsymbol{\beta}$ is known If this is the first iteration, set $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$. Otherwise, set $\hat{\boldsymbol{\beta}}$ equal to the estimator of $\boldsymbol{\beta}$ obtained from Step 2 of the previous iteration. We now estimate $\mathbf{\Gamma}$ by solving

$$\hat{\mathbf{\Gamma}} = \underset{\mathbf{\Gamma}}{\operatorname{argmin}} \{L(\hat{\boldsymbol{\beta}}, \mathbf{\Gamma})\}.$$

For each j and k, we choose our estimator of $\boldsymbol{\gamma}_{k,j}$ as

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{k,j} &= \underset{\boldsymbol{\gamma}_{k,j}}{\operatorname{argmin}}\Big\{\sum_{i=2}^{n}\big(y_{k,i} - \sum_{m=0}^{q} g_{k,j}^{(m)} x_{m,i} \\ &\quad - \sum_{m=0}^{q} \dot{g}_{k,j}^{(m)} x_{m,i}(X_{i-1} - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}\big)^2 K_h\big((X_{i-1} - X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}\big)\Big\}. \quad (4.8)\end{aligned}$$

We can write (4.8) in matrix notation as

$$\hat{\boldsymbol{\gamma}}_{k,j} = \underset{\boldsymbol{\gamma}_{k,j}}{\operatorname{argmin}}(\mathbf{y}_k - \mathcal{X}_j\boldsymbol{\gamma}_{k,j})^{\mathrm{T}} W_j(\mathbf{y}_k - \mathcal{X}_j\boldsymbol{\gamma}_{k,j})$$

where W_j and $\mathcal{X}_j$ are defined by (3.11) and (3.12) respectively. By least squares theory, the solution is given by

$$\hat{\boldsymbol{\gamma}}_{kj} = \{\mathcal{X}_j^{\mathrm{T}} W_j \mathcal{X}_j\}^{-1}\mathcal{X}_j^{\mathrm{T}} W_j \mathbf{y}_k$$

and hence we obtain $\hat{\boldsymbol{\Gamma}}_k = \{\hat{\boldsymbol{\gamma}}_{k,1}, \cdots, \hat{\boldsymbol{\gamma}}_{k,n-1}\}$ and $\hat{\boldsymbol{\Gamma}} = \{\hat{\boldsymbol{\Gamma}}_1, \cdots, \hat{\boldsymbol{\Gamma}}_{p_n}\}$.

Step 2: Estimate $\boldsymbol{\beta}$ assuming $\boldsymbol{\Gamma}$ is known Using the estimates $\hat{\boldsymbol{\Gamma}}$, from Step 1, we would like to choose our estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{(1)}$, such that

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{L(\boldsymbol{\beta}, \hat{\boldsymbol{\Gamma}})\}$$

which is equivalent to

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(1)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\Big\{\sum_{j=1}^{n-1}\sum_{i=2}^{n}\sum_{k=1}^{p_n}\{y_{k,i} - \sum_{m=0}^{q} g_{k,j}^{(m)} x_{m,i} \\ &\quad - \sum_{m=0}^{q} \dot{g}_{k,j}^{(m)} x_{m,i}(X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta}\}^2 K_h((X_{i-1} - X_j)^{\mathrm{T}}\boldsymbol{\beta})\Big\}.\end{aligned}$$

In exactly the same way as the univariate case, we approximate this minimisation using

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(1)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{j=1}^{n-1}\sum_{i=2}^{n}\sum_{k=1}^{p_n}\{y_{k,i}-\sum_{m=0}^{q} g_{k,j}^{(m)} x_{m,i} \\ &\qquad -\sum_{m=0}^{q}\dot{g}_{k,j}^{(m)} x_{m,i}(X_{i-1}-X_j)^{\mathrm{T}}\boldsymbol{\beta}\}^2 K_h((X_{i-1}-X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(0)})\}\end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{(0)}$ is the estimator for $\boldsymbol{\beta}$ used in Step 1. Rewriting the minimisation problem yields

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\sum_{j=1}^{n-1}\sum_{i=2}^{n}\sum_{k=1}^{p_n}\left\{c_{ijk}-M_{ijk}^{\mathrm{T}}\boldsymbol{\beta}\right\}^2 w_{i-1,j} \tag{4.9}$$

where

$$c_{ijk} = y_{k,i}-\sum_{m=0}^{q}\hat{g}_{k,j}^{(m)} x_{m,i}, \quad M_{ijk} = (\sum_{m=0}^{q}\hat{\dot{g}}_{k,j}^{(m)} x_{m,i})(X_{i-1}-X_j),$$

$$w_{ij} = K_h((X_i-X_j)^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{(0)}).$$

All that remains is to rewrite the triple summation in equation (4.9) in the form of a single summation of $N=(n-1)(n-1)p_n$ quantities, so that it becomes a traditional weighted least squares problem. This can be achieved by defining the $N\times 1$ matrix $\mathbf{C}$, the $N\times q$ matrix $\mathbf{M}$ and the $N\times N$ matrix $\mathbf{W}$

$$\begin{gathered}\mathbf{C} = (C_{21}^{\mathrm{T}},\cdots,C_{n1}^{\mathrm{T}},C_{22}^{\mathrm{T}},\cdots,C_{n2}^{\mathrm{T}},\cdots,C_{2,(n-1)}^{\mathrm{T}},\cdots,C_{n,(n-1)}^{\mathrm{T}})^{\mathrm{T}} \\ \mathbf{M} = (M_{21}^{\mathrm{T}},\cdots,M_{n1}^{\mathrm{T}},M_{22}^{\mathrm{T}},\cdots,M_{n2}^{\mathrm{T}},\cdots,M_{2,(n-1)}^{\mathrm{T}},\cdots,M_{n,(n-1)}^{\mathrm{T}})^{\mathrm{T}} \\ \mathbf{W} = \operatorname{diag}\{W_{21},\cdots,W_{n1},W_{22},\cdots,W_{n2},\cdots,W_{2,n-1},\cdots,W_{n,n-1}\}\end{gathered}$$

where I_{p_n} is the $p_n \times p_n$ identity matrix and where

$$C_{ij} = (c_{ij1}, \cdots, c_{ijp_n})^{\mathrm{T}}, \quad M_{ij} = (M_{ij1}, \cdots, M_{ijp_n})^{\mathrm{T}}, \quad W_{ij} = I_{p_n} w_{ij}.$$

With this notation, equation (4.9) can be written as a traditional weighted least squares problem:

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left\{ \mathbf{C}_{[i]} - \mathbf{M}_{[i]}\boldsymbol{\beta} \right\}^2 \mathbf{W}_{[i]}$$

where $\mathbf{M}_{[i]}$ and $\mathbf{C}_{[i]}$, denote the ith row of $\mathbf{M}$ and $\mathbf{C}$ respectively and where $\mathbf{W}_{[i]}$ denotes the ith diagonal entry. In matrix notation, this is equivalent to

$$\hat{\boldsymbol{\beta}}_{(1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\mathbf{C} - \mathbf{M}\boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{W} \left(\mathbf{C} - \mathbf{M}\boldsymbol{\beta}\right),$$

and the solution is given by

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{C}. \tag{4.10}$$

At this point, $\hat{\boldsymbol{\beta}}_{(1)}$ should be rescaled to satisfy the identifiability conditions $\|\boldsymbol{\beta}\| = 1$ and $\beta_1 > 0$.

As with the univariate case, care must be taken from a computational point of view to avoid exceeding the limitations of RAM because the dimensions of $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{W}$ increase at rate $O(n^2 p_n)$. For example, for a sample size of $n = 1000$ and $p_n = 100$, $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{W}$ each have approximately 10^8 rows. To deal with this problem, in the numerical studies in this thesis, (4.10) was computed using the following equations

$$\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{M} = \sum_{i=1}^{N} f_1(i) = \sum_{i=1}^{(n-1)} \sum_{j=1}^{(n-1)p_n} f_1(j + np_n(i-1)) \tag{4.11}$$

$$\mathbf{M}^{\mathrm{T}}\mathbf{W}\mathbf{C} = \sum_{i=1}^{N} f_2(i) = \sum_{i=1}^{(n-1)} \sum_{j=1}^{(n-1)p_n} f_2(j + np_n(i-1)) \tag{4.12}$$

where

$$f_1(i) = \mathbf{W}_{[i]} \begin{pmatrix} \mathbf{M}_{[i,1]}\mathbf{M}_{[i,1]} & \cdots & \mathbf{M}_{[i,1]}\mathbf{M}_{[i,q]} \\ \vdots & \ddots & \vdots \\ \mathbf{M}_{[i,q]}\mathbf{M}_{[i,1]} & \cdots\cdots & \mathbf{M}_{[i,q]}\mathbf{M}_{[i,q]} \end{pmatrix},$$

$$f_2(i) = \mathbf{W}_{[i]}\mathbf{C}_{[i]}(\mathbf{M}_{[i,1]}, \cdots, \mathbf{M}_{[i,q]})^{\mathrm{T}},$$

and $\mathbf{M}_{[i,j]}$ denotes the (i, j) entry of $\mathbf{M}$. In (4.11) and (4.12) we break the large summand into $(n-1)$ blocks of size $(n-1)p_n$ and only store one block in RAM at a time.

4.2.2 Averaging univariate estimates

In this section we introduce an alternative approach for estimating $\boldsymbol{\beta}$. We pretend, temporarily, that the true model is

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + \epsilon_{k,t}, \tag{4.13}$$

where $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,q})^{\mathrm{T}}$ and $\|\boldsymbol{\beta}_k\| = 1$, $\beta_{k,1} > 0$ for $k = 1, \cdots, p_n$. By viewing (4.13) as p_n univariate models, one can estimate $\boldsymbol{\beta}$ by estimating $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{p_n}$ using univariate methodology, and then take an average. Denoting the univariate estimator of $\boldsymbol{\beta}_k$ from Section 3.2.1 and Section 3.2.2 by $\check{\boldsymbol{\beta}}_{\mathrm{F},k}$ and $\check{\boldsymbol{\beta}}_{\mathrm{L},k}$ respectively, we introduce the following two estimators

$$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}} = \frac{1}{p_n}\sum_{k=1}^{p_n}\check{\boldsymbol{\beta}}_{\mathrm{F},k}, \quad \text{and} \quad \hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}} = \frac{1}{p_n}\sum_{k=1}^{p_n}\check{\boldsymbol{\beta}}_{\mathrm{L},k}.$$

We remark that the variance of a preliminary estimate $\check{\boldsymbol{\beta}}_k$ could be large since it only uses a portion of the information available. However, by averaging all of the p_n preliminary estimates in this way, the overall variance will be reduced. In fact, we will see that the performance of $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ are fairly similar.

4.3 Simulation study

In this section, we compare the performance of estimators $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, using a simulated example. We generate 1000 datasets from model (4.1) each with a sample size of $n = 1000$ and number of assets $p_n = 50$. We set

$$q = 4, \quad \text{and} \quad \boldsymbol{\beta} = \frac{1}{3}(1,\ 2,\ 0,\ 2)^{\mathrm{T}},$$

and for $k = 1, \cdots, p_n$ we set

$$g_k(z) = \Xi_{0,k} + 3\mathrm{exp}(-z^2), \quad a_{k,1}(z) = \Xi_{1,k} + 0.8z, \quad a_{k,2}(z) = \Xi_{2,k}$$

$$a_{k,3}(z) = \Xi_{3,k} + 1.5\mathrm{sin}(\pi z), \quad a_{k,4}(z) = \Xi_{4,k}$$

where $\Xi_{j,k}$ are some fixed parameters for $j = 0, \cdots, q$ and $k = 1, \cdots, p_n$. In order to define $\Xi_{j,k}$, we simulate them independently from a uniform distribution on $[-1, 1]$, and use these same values throughout all simulations. We generate $X_0, \cdots, X_n$ independently from a uniform distribution on $[-1, 1]^q$ and $\epsilon_{k,t}$ independently from a standard normal distribution for $k = 1, \cdots, p_n$ and $t = 1, \cdots, n$. Once both X_t and $\boldsymbol{\epsilon}_t$ have been generated, Y_t can be generated through model (4.1) for

$t = 1, \cdots, n$.

For each generated dataset, we estimate $\boldsymbol{\beta}$ using the three estimators $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$. In each case, we try the initial estimators $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ given in (3.17). In addition to this we estimate $\boldsymbol{\beta}$ using the "Oracle" estimator $\hat{\boldsymbol{\beta}}_{\mathrm{O}} = \boldsymbol{\beta}$ which uses the true $\boldsymbol{\beta}$ as the estimator.

For each estimate of $\boldsymbol{\beta}$, ($\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_{\mathrm{O}}\}$ we estimate coefficient functions $g_j(\cdot)$, $j = 0, \cdots, q$, over 101 equally spaced grid points of the following range

$$r(\hat{\boldsymbol{\beta}}) = \max(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \min(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}).$$

and denote these grid points by $u_1, \cdots, u_{101}$. We choose the bandwidth h in the same way as we did in Section 3.3.

Evaluating the performance of an estimator $\hat{\boldsymbol{\beta}}$ can be measured using $\Delta(\hat{\boldsymbol{\beta}})$ as defined in (3.19). The performance of $\hat{\mathbf{g}}(\cdot)$ and $\hat{\boldsymbol{\Phi}}(\cdot)$ are evaluated similarly using the following metric

$$\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}) = \frac{\sum_{j=0}^{q}\sum_{k=1}^{p_n}\sum_{t=1}^{100} |\hat{g}_{k,j}(u_t) - g_{k,j}(u_t)|}{\sum_{j=0}^{q}\sum_{k=1}^{p_n}\sum_{t=1}^{100} |g_{k,j}(u_t)|} \tag{4.14}$$

where

$$g_{k,j}(\cdot) = \begin{cases} a_{k,j}(\cdot) & \text{if } j = 1, \cdots, q \\ g_k(\cdot) & \text{if } j = 0. \end{cases}$$

A comparison of the expectation and standard deviation of $\Delta(\hat{\boldsymbol{\beta}})$ and $\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}})$ over the 1000 replications is given in Table 4.1. In a similar way to Chapter 3, we see that both $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ perform much better than $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ as one would expect. This is because the estimator $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ sometimes fails to converge or converges to local optima. Also, we see that the expectation and standard deviation of $\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}})$ for both $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$

and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are similar to that of the Oracle estimator $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$.

From the boxplots in Figure 10, we see that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are almost identical in performance when $\tilde{\boldsymbol{\beta}}_1$ is used. However, a key observation is that there are several occurrences when $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ performs slightly worse than $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ if $\tilde{\boldsymbol{\beta}}_2$ is used. The crucial point is that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is not sensitive at all to the choice of the initial value. We remark that the boxplots also confirm that a very large approximation error occurs if $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ is used.

The boxplots in Figure 11 also show that the coefficient functions can be estimated within a good degree of accuracy when $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are used, but less so with $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$. Typical estimated coefficient functions for these look almost identical to those presented in the previous chapter and so are omitted for brevity.

We see from Table 4.2 that $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ also outperforms $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ in terms computational time. This is because the time taken to calculate $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ on average is roughly p_n multiplied the time taken for the corresponding univariate estimation. In the real data analysis chapter of this thesis, we will calculate a new $\boldsymbol{\beta}$ for 5000 trading days and across four independent datasets. Hence this gain in computational time is very important for implementation purposes.

To conclude, we firstly see that the performance of $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ both significantly outperform $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$. Secondly, the performance $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are fairly similar, which is to be expected since they are based on similar techniques. However there is evidence to suggest that $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ is slightly more sensitive to the initial value $\tilde{\boldsymbol{\beta}}$ than $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is. Hence, the estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is the preferred choice.

Table 4.1: Comparison of estimators of $\boldsymbol{\beta}$

	$E(\Delta(\hat{\boldsymbol{\beta}}))$	$\mathrm{SD}(\Delta(\hat{\boldsymbol{\beta}}))$	$E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$	$\mathrm{SD}(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_1$	0.520	0.398	0.241	0.054
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_2$	0.673	0.437	0.259	0.060
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$	0.013	0.006	0.192	0.016
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$	0.013	0.006	0.193	0.017
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_1$	0.013	0.006	0.193	0.017
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_2$	0.015	0.018	0.193	0.017
$\hat{\boldsymbol{\beta}}_{\mathrm{O}}$	0.000	0.000	0.192	0.016

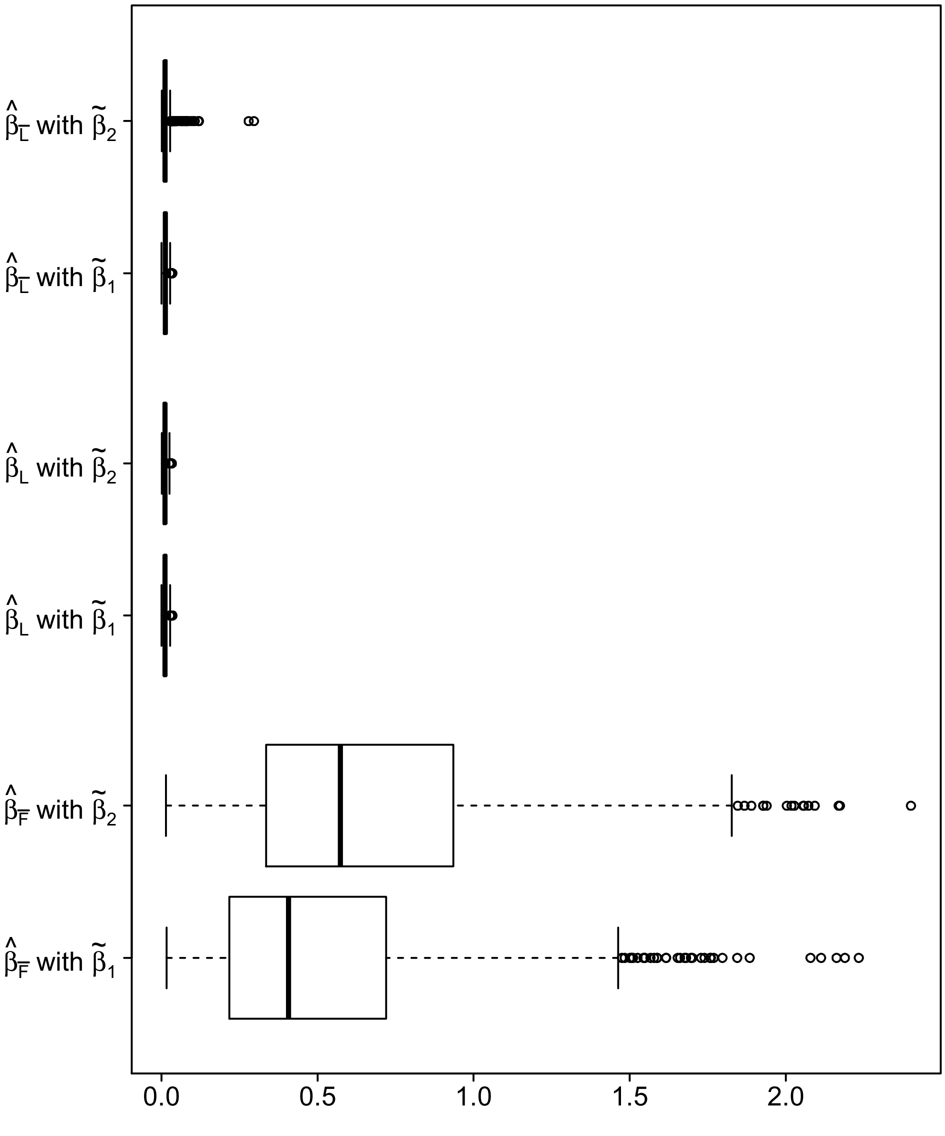
This table gives a comparison of $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ with the two initial values $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$, as well as $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$, for the simulation study in Section 4.3. The expectation and standard deviation of $\Delta(\hat{\boldsymbol{\beta}})$ and $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$ taken over 1000 replications, are given in each case.

Table 4.2: Computational time

	Mean	Standard deviation
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_1$	473.87	46.61
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ using $\tilde{\boldsymbol{\beta}}_2$	481.50	42.83
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_1$	75.59	25.44
$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ using $\tilde{\boldsymbol{\beta}}_2$	183.25	34.35
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_1$	194.88	21.38
$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ using $\tilde{\boldsymbol{\beta}}_2$	340.65	39.64

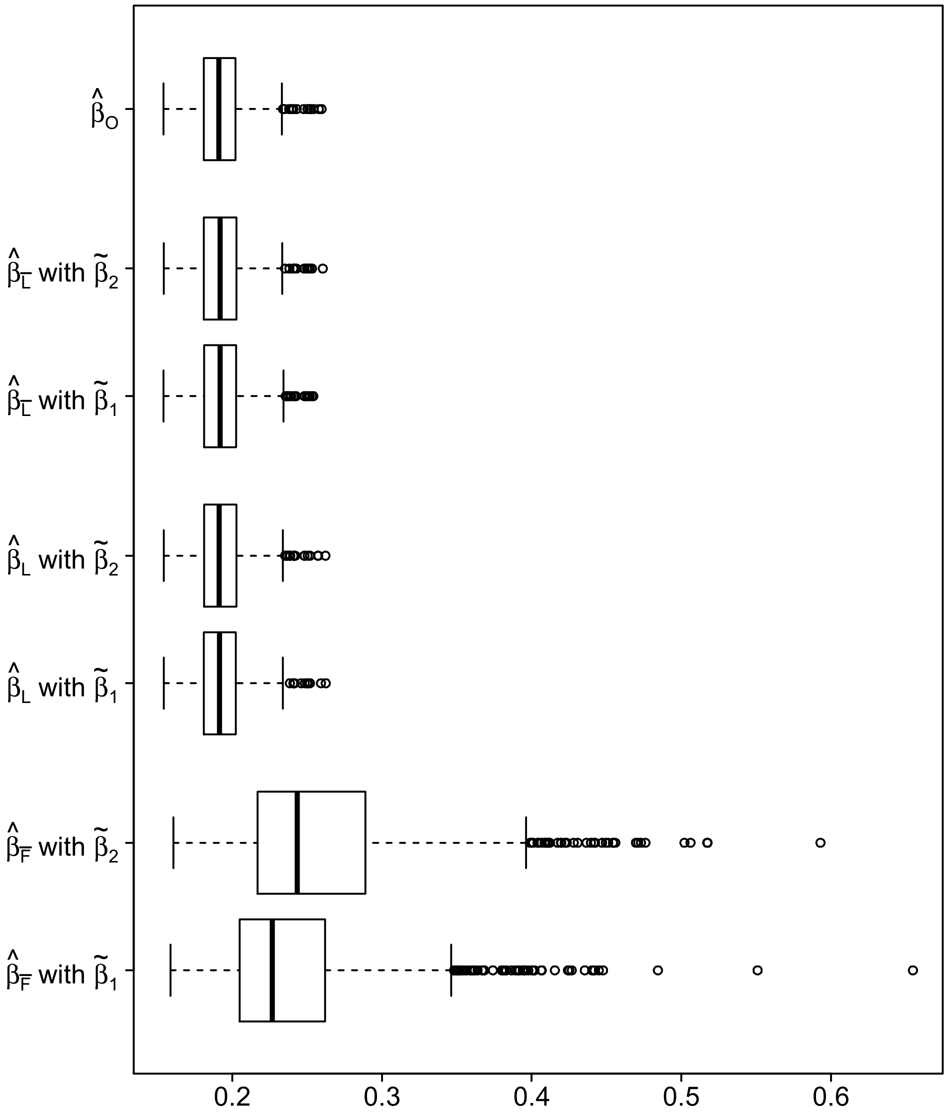
This table gives the mean and standard deviation of the computational time (seconds) to estimate $\boldsymbol{\beta}$ using $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ or $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ for a single dataset in the simulation study in Section 4.3.

Figure 10: Boxplots of $\Delta_i(\hat{\boldsymbol{\beta}})$



This figure shows boxplots of relative error metric $\Delta(\hat{\boldsymbol{\beta}})$ over 1000 replications for the simulation study in Section 4.3. A comparison of $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ with two different initial values $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ is provided.

Figure 11: Boxplots of $\Delta_i(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}})$



This figure shows boxplots of the relative error metric $\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}})$ over 1000 replications for the simulation study in Section 4.3. A comparison of $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ with two different initial values $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ is provided, along with the Oracle estimator $\hat{\boldsymbol{\beta}}_{\mathrm{O}}$.

]

5 Bandwidth selection

In this chapter, we discuss how to choose the bandwidth h used in the estimation of $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ described in Section 4.2. Since the index of the varying coefficient model is unknown until $\boldsymbol{\beta}$ has been estimated, it can be quite hard to visualise if a particular value of h is "large" or "small". Therefore, we introduce the following metric

$$h^{(\%)} = \frac{h}{r(\hat{\boldsymbol{\beta}})} \times 100 \tag{5.1}$$

where

$$r(\hat{\boldsymbol{\beta}}) = \max(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \min(X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \tag{5.2}$$

for some given estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. One can interpret $h^{(\%)}$ as the percentage of the range of estimated indices which is covered by the global bandwidth h.

It is also possible to estimate of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ using a nearest neighbour bandwidth. In this case, the size of the bandwidth at a given grid point z changes according to how many local data points are in its neighbourhood. More precisely, the k-nearest neighbour bandwidth at point z is defined by

$$h_k(z) = \min(k, \mathcal{A}), \quad \mathcal{A} = \{(z - X_1^{\mathrm{T}}\hat{\boldsymbol{\beta}},), \cdots, (z - X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}})\} \tag{5.3}$$

where $\min(k, \mathcal{A})$ is the kth smallest number of the set $\mathcal{A}$. In order to estimate $\mathbf{g}(z)$ and $\boldsymbol{\Phi}(z)$ using a nearest neighbourhood bandwidth, one simply needs to replace h with $h_k(z)$ in (4.4) and (4.5) respectively. We shall also use the metric $k^{(\%)} = k/n \times 100$ to help gauge whether a particular k is "large" or "small" relative to the sample size n.

5.1 Sensitivity to the choice of bandwidth

In this section we show that the choice of bandwidth h is not crucial for successfully estimating $\boldsymbol{\beta}$ as long as h is chosen to be within a reasonable range. However, the choice of h is more important when estimating the coefficient functions $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$. We build on the simulation study in Section 4.3 by exploring how different choices of h affect the performance of the estimation of $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$. For brevity, we only use the initial value $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_2$ in this chapter to keep the number of comparisons to a minimum. We note, however, that similar conclusions can also be made for other initial values too.

Using the same simulation settings as Section 4.3, Figure 12 shows how sensitive the expected relative error metric $E(\Delta(\hat{\boldsymbol{\beta}}))$ is to the choice of global bandwidth h, measured by $h^{(\%)}$. The crucial observation is that as long as $h^{(\%)}$ is chosen within a sensible range, such as $h^{(\%)} \in (12\%,\ 50\%)$, we can estimate $\boldsymbol{\beta}$ within a 5% relative error on average using either $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ or $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$. The same cannot be said for $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, however, which incurs a much larger approximation error on average even if the best possible bandwidth is used. These findings provide additional evidence for using estimators $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ instead of $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$. We remark that the sensitivity of $E(\Delta(\hat{\boldsymbol{\beta}}_{\mathrm{L}}))$ and $E(\Delta(\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}))$ to be fairly similar as one would expect.

In order to examine how the choice of h affects the performance of estimating $\hat{\mathbf{g}}(\cdot)$ and $\hat{\boldsymbol{\Phi}}(\cdot)$, we temporarily fix $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{O}}$ so that any estimation error incurred is solely due to the approximation error of the coefficient functions. Figure 13 shows how sensitive $E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))$ is to the choice of $h^{(\%)}$ inside the range $(12\%,\ 50\%)$. This time, we see that the choice of h is essential, since the expected relative error can be reduced from 30% to 20% with a careful choice of bandwidth. We will see in Section 5.2 that a further reduction of the relative error can be made when a nearest neighbour bandwidth is employed as opposed to a global bandwidth.

5.2 Data driven bandwidth selectors

Bandwidth selection methodology, in general, can often be computationally expensive if a grid-search approach is employed. We remark that the iterative algorithm for estimating $\boldsymbol{\beta}$ can also be computationally expensive, especially for large n and large p_n, due to the high dimensional matrices involved in each iteration. On the other hand, once $\hat{\boldsymbol{\beta}}$ is given, computing $\hat{\mathbf{g}}(z)$ and $\hat{\boldsymbol{\Phi}}(z)$ using (4.4) and (4.5) is relatively fast to compute. With these facts in mind, we introduce the following bandwidth selection approach designed to find a good balance between computational costs and quality of estimation:

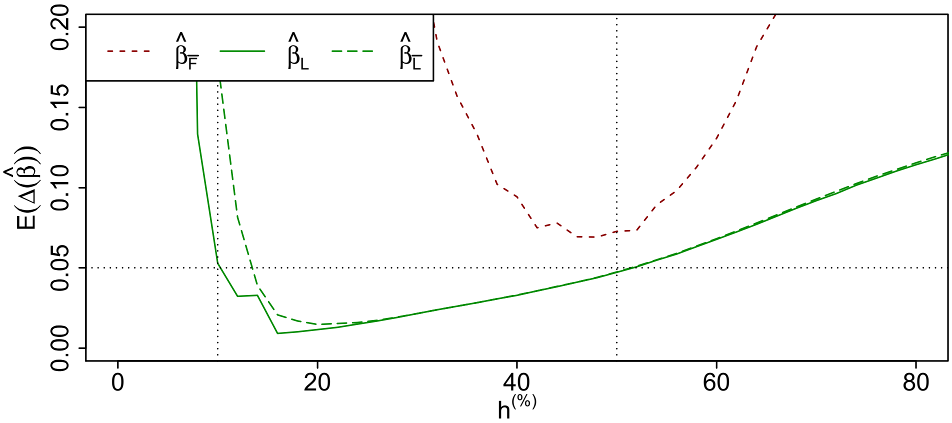
(Step 1) In the iterative algorithm used to estimate $\boldsymbol{\beta}$ (see Section 4.2.1) choose a bandwidth h equal to approximately 20% of $r(\tilde{\boldsymbol{\beta}})$ on the first iteration where $\tilde{\boldsymbol{\beta}}$ is the arbitrary initial value. Update h on all subsequent iterations by choosing h equal to approximately 20% of $r(\hat{\boldsymbol{\beta}}_{(0)})$ where $\hat{\boldsymbol{\beta}}_{(0)}$ is the most recent estimate.

(Step 2) After the algorithm has converged, and an estimate for $\boldsymbol{\beta}$ is obtained, we estimate $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ using a data driven bandwidth selector $\hat{h}_2$ in place of h in (4.4) and (4.5).

The subject of this section is to explore various data-driven methods of choosing $\hat{h}_2$ and compare their performance. In Step 2, $\hat{h}_2$ can either be a global bandwidth or a nearest neighbour bandwidth.

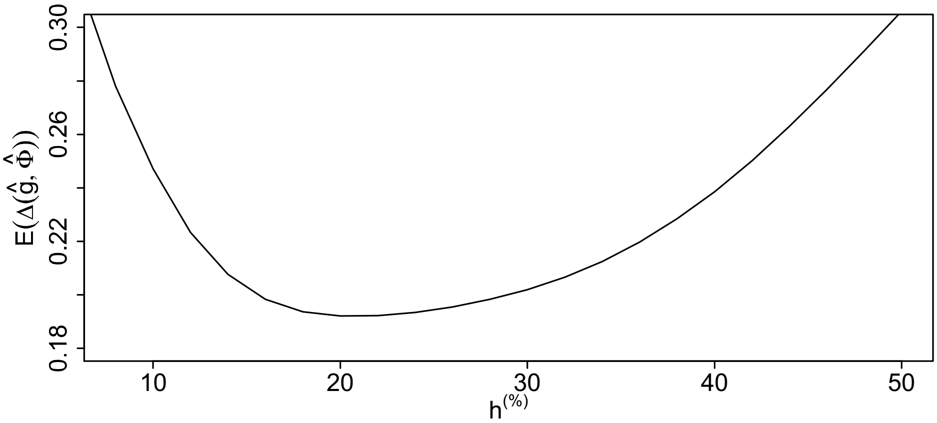
The choice of a global bandwidth $\hat{h}_2$, or $\hat{k}$ in the case of a nearest neighbour bandwidth, controls the trade off between bias and variance. On the one hand, choosing a bandwidth which is too small would result

Figure 12: Sensitivity of $E(\Delta(\hat{\boldsymbol{\beta}}))$ to the choice of h



Using the same simulation settings from Section 4.3, this figure shows $E(\Delta(\hat{\boldsymbol{\beta}}))$ *vs* $h^{(\%)}$ *for* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ *and* $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ *using* $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_2$. *The dotted vertical lines represent a reasonable region to choose* $h^{(\%)}$ *whereby* $E(\Delta(\hat{\boldsymbol{\beta}})) < 0.05$ *for* $h^{(\%)} \in (12, 50)$.

Figure 13: Sensitivity of $E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))$ to the choice of h



Using the same simulation settings from Section 4.3, this figure shows $E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))$ *vs* $h^{(\%)}$ *for* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{O}}$. *Unlike* $\hat{\boldsymbol{\beta}}$, *we see that* $\hat{\mathbf{g}}(\cdot)$ *and* $\hat{\boldsymbol{\Phi}}(\cdot)$ *require a careful choice of bandwidth to reduce estimation error as much as possible.*

in estimators for $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$ with small bias but large variance. On the other hand, choosing a bandwidth which is too large would result in estimators for $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$ with small variance but large bias. Both of these cases would result in large approximation error which is not desirable. Therefore, in order to carefully choose bandwidth h_2, we require an information criterion $I(h_2)$ which, once minimised with respect to h_2, aims to find a trade-off between bias and variance.

With this in mind, we explore three information criteria based on Akaike information criterion (AIC), Bayesian information criterion (BIC) and cross validation (CV). Each of these criteria can be used for either global bandwidth selection or nearest neighbour bandwidth selection. For implementation purposes, to minimise $I(h_2)$ with respect to h_2, we employ the following grid search approach:

(Step 1) Choose a number of candidate bandwidths. For the case of global bandwidths, we recommend $h^{(\%)} \in \mathcal{H}$ where $\mathcal{H} = \{1\%,\ 2\%, \cdots, 100\%\}$. For the case of nearest neighhbour bandwidths, we recommend $k \in \mathcal{K}$ where $\mathcal{K} = \{1, \cdots, n\}$.

(Step 2) For the case of global bandwidths, compute $I(h^{(\%)})$ for each $h^{(\%)} \in \mathcal{H}$ and choose $\hat{h}^{(\%)} = \underset{h^{(\%)} \in \mathcal{H}}{\text{argmin}}\{I(h^{(\%)})\}$. Similarly for the case of nearest neighour bandwidths, compute $I(k)$ for each $k \in \mathcal{K}$ and choose $\hat{k} = \underset{k \in \mathcal{K}}{\text{argmin}}\{I(k)\}$.

We remark that this grid search approach can implemented using parallel computing to significantly speed up the computation. Furthermore, subsets of $\mathcal{H}$ and $\mathcal{K}$ can be chosen to speed up computation if necessary, at the expense of approximating the minimisation.

5.2.1 Cross validation

In this section, we introduce an information criterion which can be used for bandwidth selection based on cross validation. Assume that an estimate $\hat{\boldsymbol{\beta}}$ has been obtained using one of the methods in Chapter 4. We define the cross validation statistic at time t by

$$\sum_{t=n-\nu}^{n} \left\| Y_t - \hat{\mathbf{g}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\Phi}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})X_t \right\| \tag{5.4}$$

where $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are the respective estimates of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ based on $(X_l^{\mathrm{T}}, Y_l^{\mathrm{T}})$, $l = 1, \cdots, t-1$, and where ν is a look-back integer such that $\nu < n-1$. We denote (5.4) by $\mathrm{CV}(h)$ when $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are estimated with a global bandwidth h, and by $\mathrm{CV}(k)$ when $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are estimated with a nearest neighbour bandwidth $h_k(\cdot)$.

Hence, we define the following global and nearest neighbour bandwidth selectors:

$$\hat{h}_{\mathrm{CV}} = \underset{h}{\mathrm{argmin}}\big\{\mathrm{CV}(h)\big\}, \quad \hat{k}_{\mathrm{CV}} = \underset{k}{\mathrm{argmin}}\big\{\mathrm{CV}(k)\big\}.$$

We define $\nu^{(\%)} = 100\nu/n$ as the percentage of the data for which we use to train our tuning parameter search on. In Section 5.3 we see that both estimators $\hat{h}_{\mathrm{CV}}$ and $\hat{k}_{\mathrm{CV}}$ are not very sensitive to the choice of $\nu^{(\%)}$ when chosen to be in a reasonable range.

5.2.2 AIC and BIC

Stimulated by Fan et al. (2003), in this section we construct AIC and BIC statistics used as information criterion for bandwidth selection. Assume that an estimate $\hat{\boldsymbol{\beta}}$ has been obtained using one of the methods in Chapter 4. We start by defining the (locally weighted) residual sum of squares metric

$$\begin{aligned}\mathrm{RSS}_k(z,h) =& \sum_{t=2}^{n}\big(y_{k,t} - g_k(z) - X_t^{\mathrm{T}}\mathbf{a}_k(z) - (\dot{g}_k(z) \\ &+ X_t^{\mathrm{T}}\dot{\mathbf{a}}_k(z))(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z)\big)^2 K_h(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}} - z), \qquad (5.5)\end{aligned}$$

for a given grid point z and for $k = 1, \cdots, p_n$. We also define the (local) degrees of freedom, $m(z,h)$, and the (local) number of observations, $n(z,h)$, by

$$m(z,h) = n(z,h) - p(z,h), \qquad n(z,h) = \mathrm{tr}\{W(z)\}$$

where

$$p(z,h) = \mathrm{tr}\{(\mathcal{X}(z)W(z)\mathcal{X}(z))^{-1}\mathcal{X}^{\mathrm{T}}(z)W^2(z)\mathcal{X}(z)\}$$

represents the (local) number of parameters. Here, $W(z)$ and $\mathcal{X}(z)$ are defined in (3.4) and (3.5) respectively. Note that h can either be a global bandwidth or a nearest neighbour bandwidth $h_k(z)$ depending on k. Using the above formulae, the local AIC and the local BIC at grid point z using bandwidth h may be defined as

$$\mathrm{AIC}_k(z,h) = \log\big(\mathrm{RSS}_k(z,h)/m(z,h)\big) + 2p(z,h)/n(z,h)$$

and

$$\mathrm{BIC}_k(z,h) = \log\big(\mathrm{RSS}_k(z,h)/m(z,h)\big) + \log(n(z,h))\; p(z,h)/n(z,h)$$

for $k = 1, \cdots, p_n$. Hence we define the following two global bandwidth selectors:

$$\hat{h}_{\text{AIC}} = \operatorname*{argmin}_h \{ \frac{1}{N_{\text{grid}} p_n} \sum_{k=1}^{p_n} \sum_{i=1}^{N_{\text{grid}}} \text{AIC}_j(z_i, h) \} \tag{5.6}$$

$$\hat{h}_{\text{BIC}} = \operatorname*{argmin}_h \{ \frac{1}{N_{\text{grid}} p_n} \sum_{k=1}^{p_n} \sum_{i=1}^{N_{\text{grid}}} \text{BIC}_j(z_i, h) \} \tag{5.7}$$

and the following two nearest neighbour bandwidth selectors:

$$\hat{k}_{\text{AIC}} = \operatorname*{argmin}_k \{ \frac{1}{N_{\text{grid}} p_n} \sum_{j=1}^{p_n} \sum_{i=1}^{N_{\text{grid}}} \text{AIC}_j(z_i, h_k(z_i)) \} \tag{5.8}$$

$$\hat{k}_{\text{BIC}} = \operatorname*{argmin}_k \{ \frac{1}{N_{\text{grid}} p_n} \sum_{j=1}^{p_n} \sum_{i=1}^{N_{\text{grid}}} \text{BIC}_j(z_i, h_k(z_i)) \} \tag{5.9}$$

where $z_1, \cdots, z_{N_{\text{grid}}}$ denote N_{grid} equally spaced grid points between $\max(X_1^{\text{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\text{T}}\hat{\boldsymbol{\beta}})$ and $\min(X_1^{\text{T}}\hat{\boldsymbol{\beta}}, \cdots, X_n^{\text{T}}\hat{\boldsymbol{\beta}})$. As explored by Fan et al. (2003), a further extension is to add a weight function to help reduce the effects at the boundary.

The reason we introduce the above AIC and BIC metrics is so that we can provide an interesting comparison with the cross validation approach. However, we will see evidence in our simulation study that cross validation using a nearest neighbour bandwidth is the preferred (and proposed) choice, and that we do not recommend using $\hat{h}_{\text{AIC}}$, $\hat{h}_{\text{BIC}}$, $\hat{k}_{\text{AIC}}$ or $\hat{k}_{\text{BIC}}$ when applied to real data.

5.3 Simulation study

In this section we compare the performance the global bandwidth selectors ($\hat{h}_{\mathrm{AIC}}$, $\hat{h}_{\mathrm{BIC}}$, and $\hat{h}_{\mathrm{CV}}$) and the nearest neighbour bandwidth selectors ($\hat{k}_{\mathrm{AIC}}$, $\hat{k}_{\mathrm{BIC}}$, and $\hat{k}_{\mathrm{CV}}$), described in Section 5.2, using a simulated example. From model (4.1), we generate 1000 datasets using $\{n = 1000,\ p_n = 50\}$ and another 1000 datasets using $\{n = 2000,\ p_n = 50\}$. This will allow us to see how the performance of the bandwidth selectors is affected by an increase in sample size. We generate $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ using the identical model structure and assumptions to Section 4.3.

In order to keep the number of comparisons to a minimum, we shall only estimate $\boldsymbol{\beta}$ using the proposed estimator $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$. We use the same initial value $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_2$ which was used in Section 4.3. Some other data analysis, not presented in this thesis for brevity, shows that same conclusions can be made for other initial values whenever $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ is used.

For each generated dataset, we estimate $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ using the global bandwidth selectors $\hat{h}_{\mathrm{AIC}}$, $\hat{h}_{\mathrm{BIC}}$ and $\hat{h}_{\mathrm{CV}}$ and nearest neighbour bandwidth selectors $\hat{k}_{\mathrm{AIC}}$, $\hat{k}_{\mathrm{BIC}}$ and $\hat{k}_{\mathrm{CV}}$. We initially set $\nu^{(\%)} = 90\%$ for cross validation selectors, and later show that this choice is arbitrary.

We compare the performance of these bandwidth selectors with the *Oracle global bandwidth* $\hat{h}_{\mathrm{O}}$ and the *Oracle nearest neighbour bandwidth* $\hat{k}_{\mathrm{O}}$ defined respectively by

$$\hat{h}_{\mathrm{O}} = \underset{h}{\operatorname{argmin}}\big\{E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))\big\} \tag{5.10}$$

and

$$\hat{k}_{\mathrm{O}} = \underset{k}{\operatorname{argmin}}\big\{E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))\big\}. \tag{5.11}$$

We can interpret $\hat{h}_{\mathrm{O}}$ as the tuning parameter which minimises the quantity $E(\Delta(\hat{\mathbf{g}}, \hat{\boldsymbol{\Phi}}))$ where a global bandwidth is used to estimate

$\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$. The interpretation for $\hat{k}_{\text{O}}$ is the same, except a nearest neighbour bandwidth is used to estimate $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$. The purpose of introducing $\hat{h}_{\text{O}}$ and $\hat{k}_{\text{O}}$ is to aid as a benchmark to compare the methods introduced in Section 5.2. One should note, however, that $\hat{h}_{\text{O}}$ and $\hat{k}_{\text{O}}$ can only be calculated in a simulation study since in practice we do not know the quantity $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$.

In Figure 14, we plot the quantity $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ against $h_2^{(\%)}$ where $\hat{\mathbf{g}}(\cdot)$ and $\hat{\mathbf{\Phi}}(\cdot)$ are estimated using a global bandwidth. Similarly in Figure 15 we plot $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ against $k^{(\%)}$ where $\hat{\mathbf{g}}(\cdot)$ and $\hat{\mathbf{\Phi}}(\cdot)$ are estimated using a nearest neighbour bandwidth. We do this for both sample sizes $n = \{1000,\ 2000\}$, and read-off the Oracle global bandwidths $\hat{h}_{\text{O}}^{(\%)} = \{21.0\%,\ 16.0\%\}$ and the Oracle nearest neighbour bandwidths $\hat{k}_{\text{O}}^{(\%)} = \{30.0\%,\ 27.5\%\}$ which correspond to the the minimisor of $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$. As one expects, we see that as the sample size increases, the values of $\hat{h}_{\text{O}}^{(\%)}$ and $\hat{k}_{\text{O}}^{(\%)}$ decrease, and the expected errors also decrease. An important point is that a much smaller relative error can be achieved on average using a nearest neighbour bandwidth instead of using a global bandwidth. This fact will be verified again later in this section.

An initial comparison of the performance of the six data-driven bandwidth selectors is given in Table 5.1 for $n = 1000$ and Table 5.2 for $n = 2000$. A vital observation is that $\hat{k}_{\text{CV}}$ performs better than the other five bandwidth selectors with $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})) = 0.15$ and 0.12 for $n = 1000$ and $n = 2000$ respectively. We see that, on average, $\hat{k}_{\text{CV}}^{(\%)}$ only slightly under-smooths relative to $\hat{k}_{\text{O}}^{(\%)}$, with $E[\hat{k}_{\text{CV}}^{(\%)}] = 29.35\%$ and 25.84%. This is also illustrated in Figure 18 and Figure 19 which show a kernel density estimate of how $\hat{k}_{\text{CV}}^{(\%)}$ is distributed relative to $\hat{k}_{\text{O}}$. We see that the mass is distributed fairly evenly around $\hat{k}_{\text{O}}^{(\%)}$ with only a very slight downwards bias.

The global bandwidth selectors $\hat{h}_{\text{AIC}}$, $\hat{h}_{\text{BIC}}$ and $\hat{h}_{\text{CV}}$ perform worse than $\hat{k}_{\mathrm{CV}}$ since the quantity $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ is larger in each case. To see why this is, it is useful to look at the boxplots of selected global bandwidths across the 1000 replications given in Figure 16 and Figure 17. We can see that $\hat{h}_{\mathrm{AIC}}$ and $\hat{h}_{\mathrm{BIC}}$ notably over smooths whereas $\hat{h}_{\mathrm{CV}}$ slightly under smooths when compared to $\hat{h}_{\mathrm{O}}^{(\%)}$.

The nearest neighbour bandwidths $\hat{k}_{\mathrm{AIC}}$ and $\hat{k}_{\mathrm{BIC}}$ have extremely poor performance, and incur a strikingly large value of $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ due to excessive over smoothing. This evidence shows that one should not select a nearest neighbour bandwidth using the AIC and BIC statistics constructed by (5.8) and (5.9). However, as a suggestion for future work, one may try an an alternative construction of the AIC and BIC statistic (see for example Cheng et al., 2009).

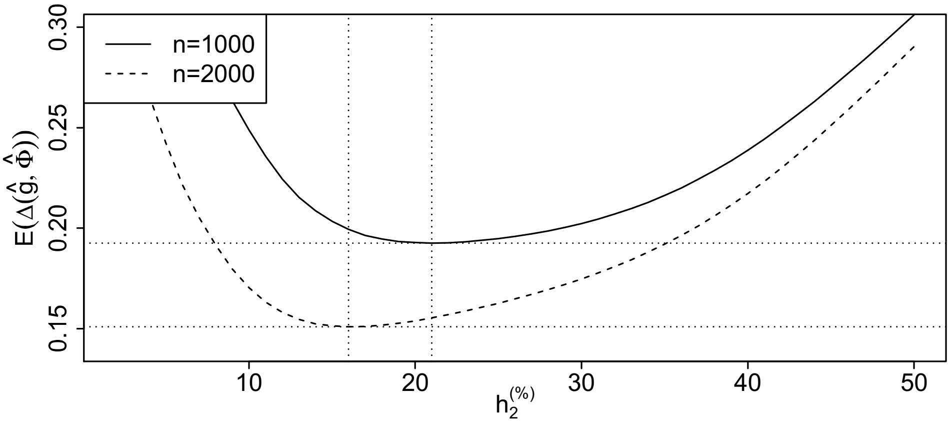
For the approaches which use cross validation it is important to see whether different values of $\nu^{(\%)}$ significantly affect the quality of estimation. One may argue that if the estimator $\hat{k}_{\mathrm{CV}}$ is highly sensitive to the choice of $\nu^{(\%)}$, then its strong performance may simply be due to chance alone. Hence, in order to address this concern, we repeated the above experiment using $\nu^{(\%)} = 60$, 70, 80 and 90. The results, displayed in Figure 24 and Figure 25, show that the estimation performance is not sensitive to the choice of $\nu^{(\%)}$. We remark however, that there is a significant increase in computation time as one chooses smaller values of $\nu^{(\%)}$. It is for this reason why fix $\nu^{(\%)} = 90$ for the rest of this thesis, however this choice is not crucial.

Another useful way to compare the six bandwidth selectors is given in Figure 20 and Figure 21 which show boxplots of the relative errors $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$ over the 1000 replications for $n = 1000$ and $n = 2000$ respectively. These plots agree with the previous analysis, and emphasise the most crucial point that the preferred way of estimating $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$ is using $\hat{k}_{\mathrm{CV}}$. It is easy to see that $\hat{k}_{\mathrm{CV}}$ also has a similar performance to that of the Oracle estimator $\hat{k}_{\mathrm{O}}$, and that as the sample size increases the approximation errors clearly reduce.

One key reason why $\hat{k}_{\text{CV}}$ outperforms the other five bandwidth selectors, in particular its closest rival $\hat{h}_{\text{CV}}$, can be illustrated in Figure 22 and Figure 23 which show plots of typical estimated coefficient functions for $g_1(\cdot)$ and $a_{1,3}(\cdot)$ respectively. We see that when $\hat{k}_{\text{CV}}$ is used, the typical estimates closely approximate the true functions at all grid points. Although $\hat{h}_{\text{CV}}$ can closely approximate the true functions at central grid points, it suffers from severe estimation error at the boundary. The reason why $\hat{k}_{\text{CV}}^{(\%)}$ is so successful is because if there are very few data points at the boundary, it will pick a slightly larger bandwidth to help reduce the variance. From these plots, it is possible to see that this is how $\hat{k}_{\text{CV}}$ outperforms $\hat{h}_{\text{CV}}$ overall. Finally, it is clear from Figure 22 and Figure 23 that both $\hat{k}_{\text{AIC}}$ and $\hat{k}_{\text{BIC}}$ fail to approximate the true function due to their excessive over smoothing.

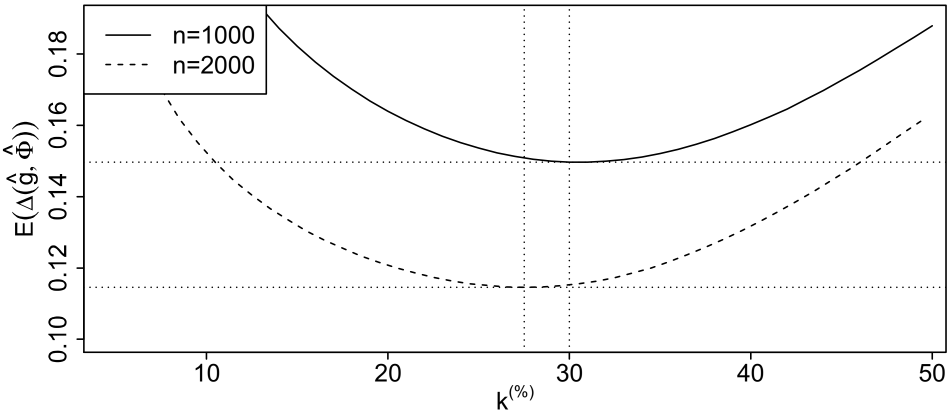
To conclude, we firstly propose to estimate $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$ using a nearest neighbour bandwidth selected by cross validation. Secondly, we have seen that $\hat{k}_{\text{CV}}$ outperforms other candidate bandwidth selectors including those using AIC and BIC statistics related to Fan et al. (2003). Finally, we have provided evidence that the performance of $\hat{k}_{\text{CV}}$ is fairly similar to $\hat{k}_{\text{O}}$ in this simulated example.

Figure 14: Visualisation for finding $\hat{h}_{\mathrm{O}}$



This figure shows $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *vs* $h_2^{(\%)}$ *when* $\hat{\mathbf{g}}(\cdot)$ *and* $\hat{\mathbf{\Phi}}(\cdot)$ *are estimated using a global bandwidth in the simulation study in Section 5.3. Vertical dotted lines correspond to* $\hat{h}_{\mathrm{O}}$ *and horizontal dotted lines correspond to the value of* $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *attained when* $\hat{h}_{\mathrm{O}}$ *is used to estimate* $\mathbf{g}(\cdot)$ *and* $\mathbf{\Phi}(\cdot)$*.*

Figure 15: Visualisation for finding $\hat{k}_{\mathrm{O}}$



This figure shows $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *vs* $k^{(\%)}$ *when* $\hat{\mathbf{g}}(\cdot)$ *and* $\hat{\mathbf{\Phi}}(\cdot)$ *are estimated using a nearest neighbour bandwidth in the simulation study in Section 5.3. Vertical dotted lines correspond to* $\hat{k}_{\mathrm{O}}$ *and horizontal dotted lines correspond to the value of* $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *attained when* $\hat{k}_{\mathrm{O}}$ *is used to estimate* $\mathbf{g}(\cdot)$ *and* $\mathbf{\Phi}(\cdot)$*.*

Table 5.1: Comparison of bandwidth selectors $n = 1000$

	$E[h_2^{(\%)}]$	$\mathrm{SD}[h_2^{(\%)}]$	$E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$	$\mathrm{SD}(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$
$\hat{h}_{\mathrm{AIC}}$	33.35	1.44	0.21	0.01
$\hat{h}_{\mathrm{BIC}}$	35.70	1.39	0.22	0.01
$\hat{h}_{\mathrm{CV}}$	14.00	2.39	0.21	0.04
	$E[k^{(\%)}]$	$\mathrm{SD}[k^{(\%)}]$	$E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$	$\mathrm{SD}(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$
$\hat{k}_{\mathrm{AIC}}$	98.75	0.95	0.47	0.03
$\hat{k}_{\mathrm{BIC}}$	98.87	0.76	0.47	0.03
$\hat{k}_{\mathrm{CV}}$	29.35	2.72	0.15	0.01

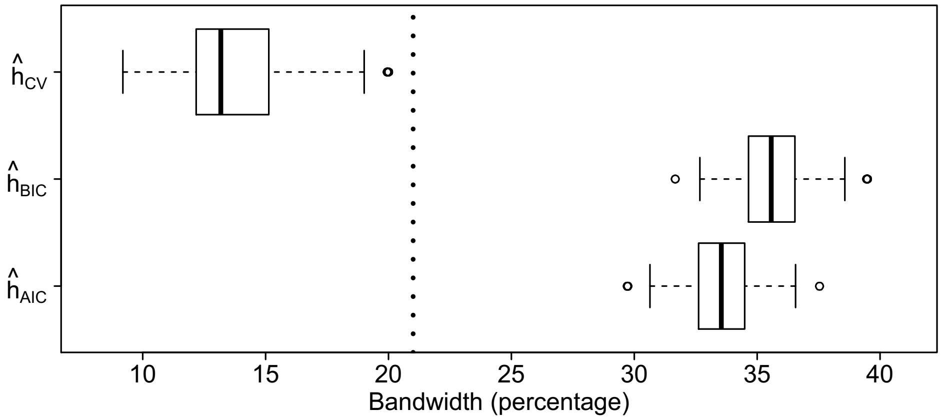
This table shows expectation and standard deviation of the estimated bandwidths along with $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *and* $\mathrm{SD}(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$*, for the simulation study in Section 5.3 with a sample size* $n = 1000$.

Table 5.2: Comparison of bandwidth selectors for $n = 2000$

	$E[h_2^{(\%)}]$	$\mathrm{SD}[h_2^{(\%)}]$	$E[\Delta(\hat{\mathbf{g}}, \hat{A})]$	$\mathrm{SD}[\Delta(\hat{\mathbf{g}}, \hat{A})]$
$\hat{h}_{\mathrm{AIC}}$	30.99	1.36	0.18	0.01
$\hat{h}_{\mathrm{BIC}}$	33.12	1.26	0.19	0.01
$\hat{h}_{\mathrm{CV}}$	11.07	1.41	0.17	0.03
	$E[k^{(\%)}]$	$\mathrm{SD}[k^{(\%)}]$	$E[\Delta(\hat{\mathbf{g}}, \hat{A})]$	$\mathrm{SD}[\Delta(\hat{\mathbf{g}}, \hat{A})]$
$\hat{k}_{\mathrm{AIC}}$	98.79	0.63	0.46	0.02
$\hat{k}_{\mathrm{BIC}}$	98.85	0.59	0.46	0.02
$\hat{k}_{\mathrm{CV}}$	25.84	2.29	0.12	0.01

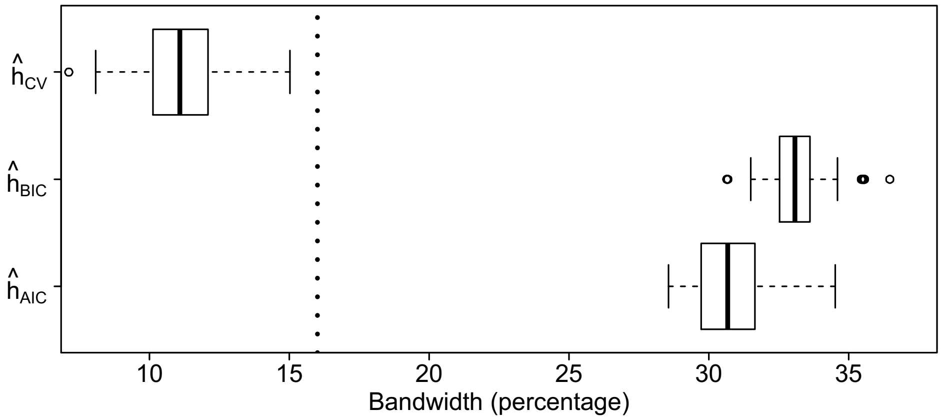
This table shows expectation and standard deviation of the estimated bandwidths along with $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *and* $\mathrm{SD}(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$*, for the simulation study in Section 5.3 with a sample size* $n = 2000$.

Figure 16: Selected global bandwidths for $n = 1000$



For the simulation study in Section 5.3, this figure shows boxplots of selected $\hat{h}^{(\%)}_{\text{AIC}}$, $\hat{h}^{(\%)}_{\text{BIC}}$, $\hat{h}^{(\%)}_{\text{CV}}$ *over 1000 replications for* $n = 1000$. *The dotted line corresponds to* $\hat{h}^{(\%)}_{\text{O}} = 21\%$.

Figure 17: Selected global bandwidths for $n = 2000$



For the simulation study in Section 5.3, this figure shows boxplots of selected $\hat{h}^{(\%)}_{\text{AIC}}$, $\hat{h}^{(\%)}_{\text{BIC}}$, $\hat{h}^{(\%)}_{\text{CV}}$ *over 1000 replications for* $n = 2000$. *The dotted line corresponds to* $\hat{h}^{(\%)}_{\text{O}} = 16\%$.

Figure 18: Density estimates of $\hat{k}_{\text{CV}}$ ($n = 1000$)

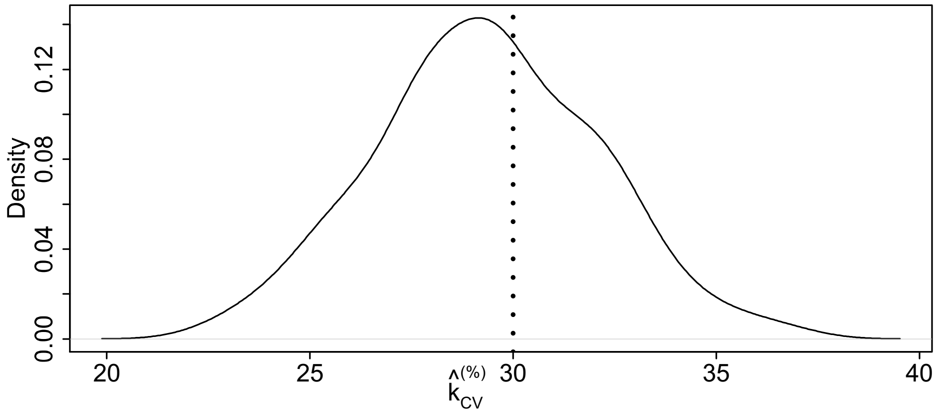
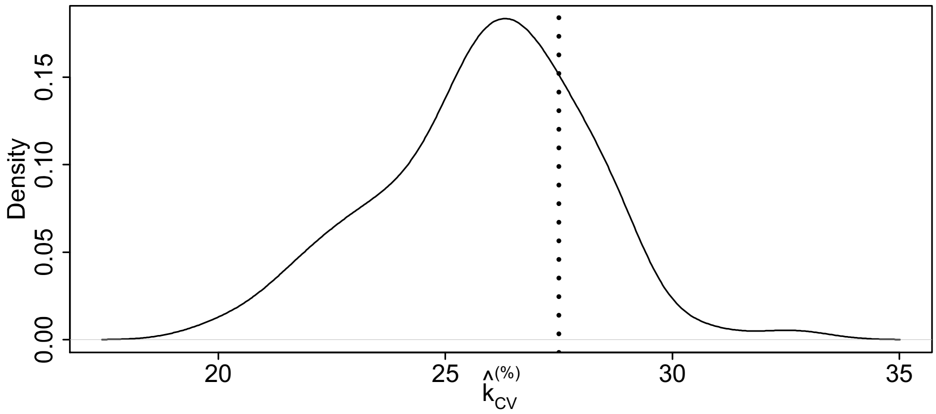
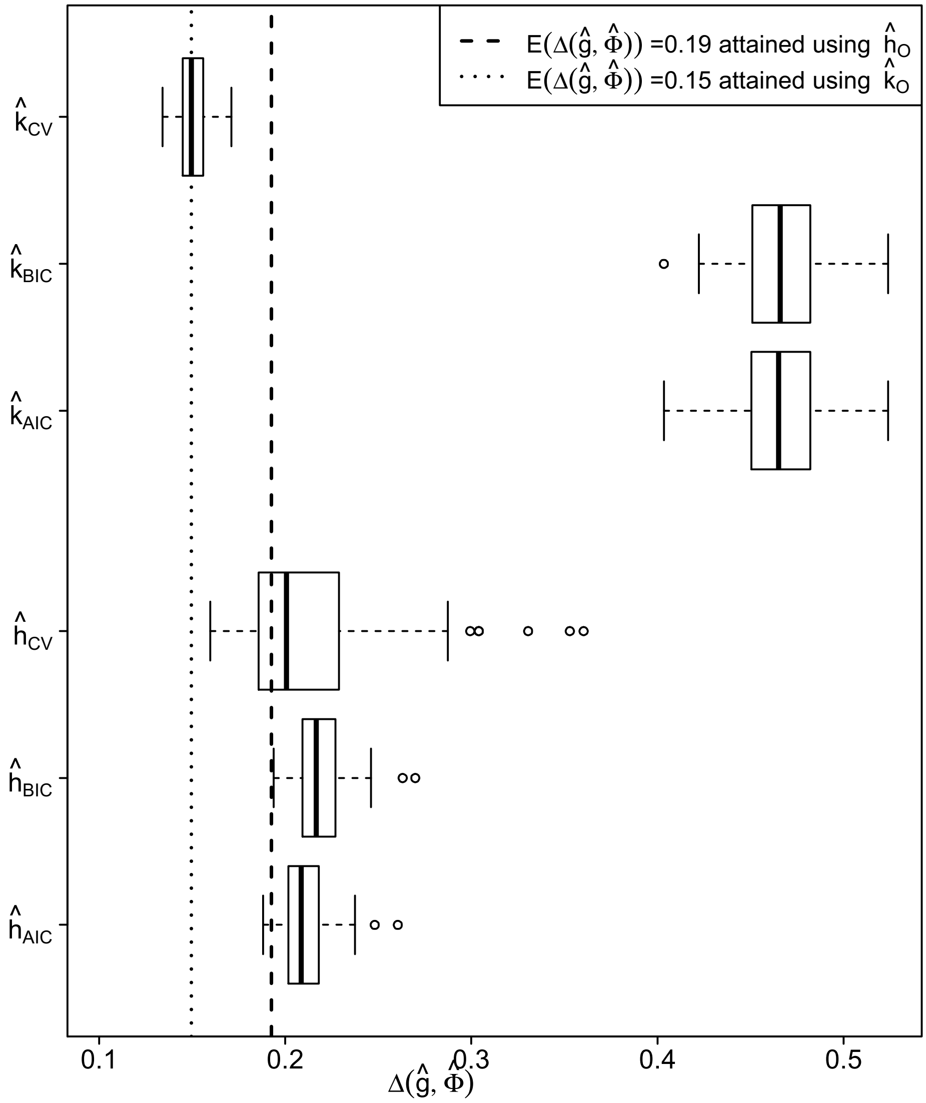


Figure 19: Density estimates of $\hat{k}_{\text{CV}}$ ($n = 2000$)



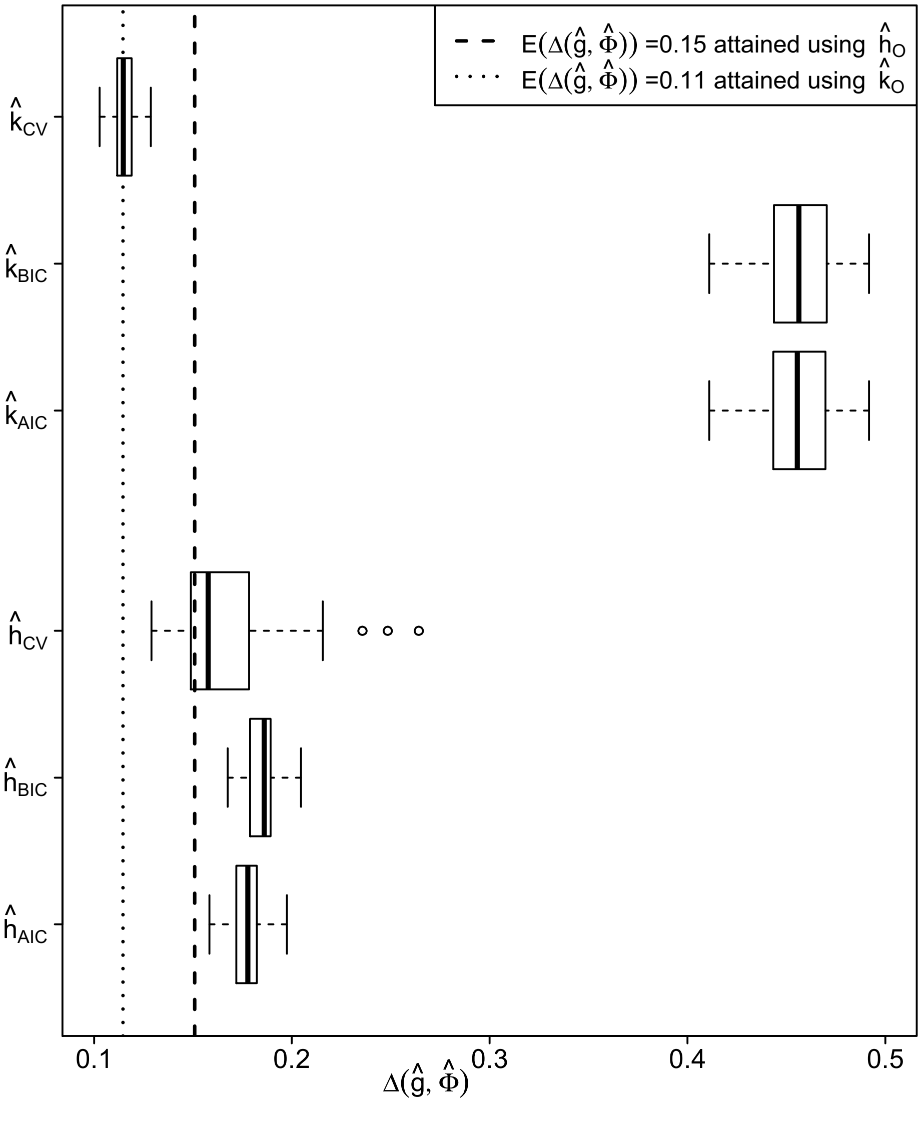
For the simulation study in Section 5.3, Figure 18 and Figure 19 show kernel density estimates of selected $\hat{k}_{\text{CV}}$ *for* $n = 1000$ *and* $n = 2000$ *respectively. The vertical dotted lines correspond to the Oracle estimator* $\hat{k}_{\text{O}}^{(\%)}$*. To produce the kernel density estimate, bandwidths were chosen using Silverman's 'rule of thumb' bandwidth selector (Silverman, 1986) which is default in R.*

Figure 20: Comparison of bandwidth selectors ($n = 1000$)



For the simulation study in Section 5.3, this figure shows boxplots of $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$ for $\hat{h}_{\mathrm{AIC}}$, $\hat{h}_{\mathrm{BIC}}$, $\hat{h}_{\mathrm{CV}}$, $\hat{k}_{\mathrm{AIC}}$, $\hat{k}_{\mathrm{BIC}}$, and $\hat{k}_{\mathrm{CV}}$ for sample size $n = 1000$. Vertical dashed and dotted lines correspond to the benchmark values $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ attained by $\hat{h}_{\mathrm{O}}^{(\%)}$ and $\hat{k}_{\mathrm{O}}^{(\%)}$ respectively.

Figure 21: Comparison of bandwidth selectors ($n = 2000$)



For the simulation study in Section 5.3, this figure shows boxplots of $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$ *for* $\hat{h}_{\text{AIC}}$, $\hat{h}_{\text{BIC}}$, $\hat{h}_{\text{CV}}$, $\hat{k}_{\text{AIC}}$, $\hat{k}_{\text{BIC}}$, *and* $\hat{k}_{\text{CV}}$ *for sample size* $n = 2000$. *Vertical dashed and dotted lines correspond to the benchmark values* $E(\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}}))$ *attained by* $\hat{h}_{\text{O}}^{(\%)}$ *and* $\hat{k}_{\text{O}}^{(\%)}$ *respectively.*

Figure 22: Typical estimates of $g_1(\cdot)$

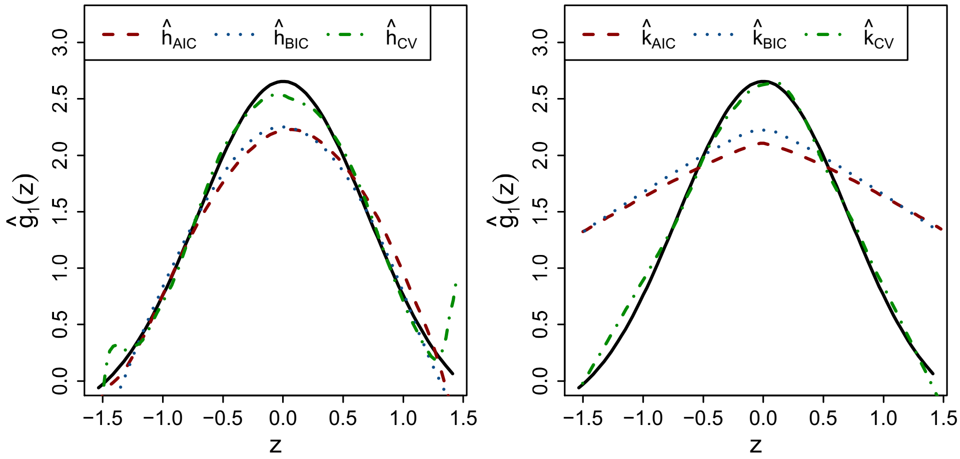
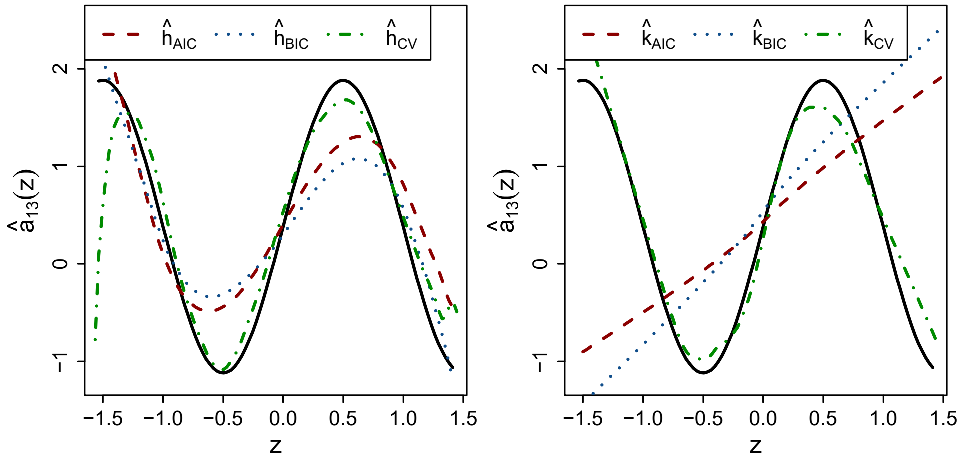
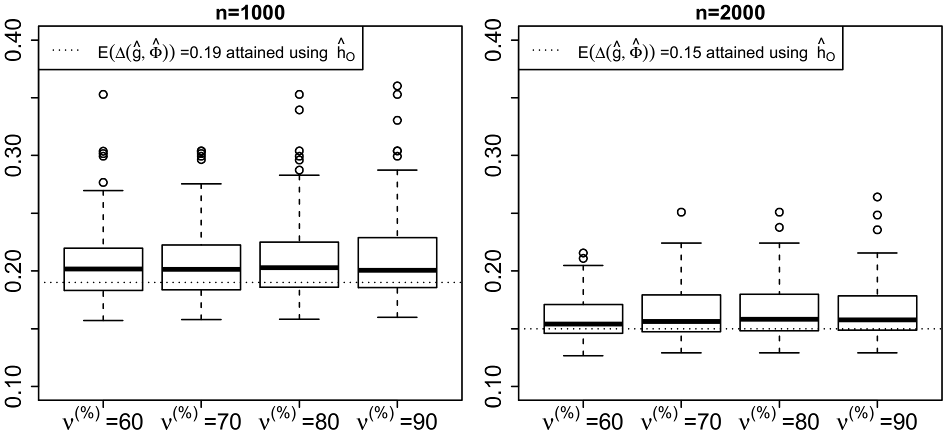


Figure 23: Typical estimates of $a_{1,3}(\cdot)$



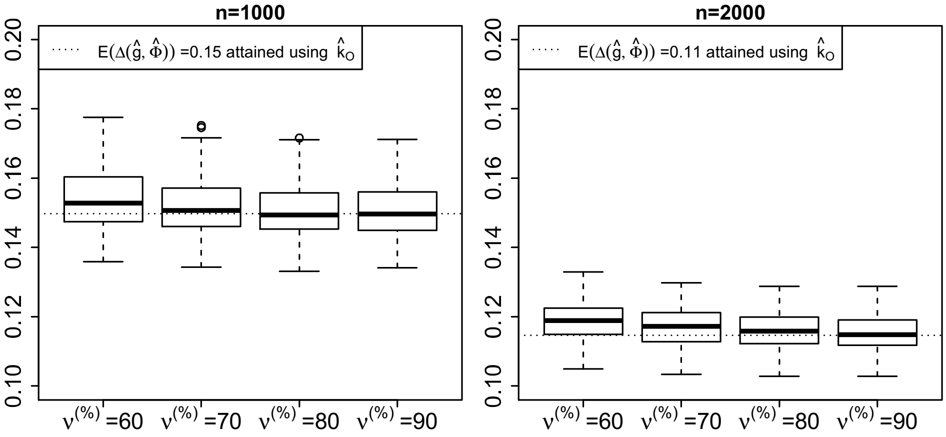
For the simulation study in Section 5.3, Figure 22 and Figure 23 show typical estimates of $g_1(\cdot)$ and $a_{1,3}(\cdot)$ respectively. The typical estimates correspond to seed with the median value of $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$ over 1000 simulations. On the left are global bandwidth selectors $\hat{h}_{\mathrm{AIC}}$, $\hat{h}_{\mathrm{BIC}}$, $\hat{h}_{\mathrm{CV}}$, and on the right are the nearest neighbour bandwidth selectors $\hat{k}_{\mathrm{AIC}}$, $\hat{k}_{\mathrm{BIC}}$, $\hat{k}_{\mathrm{CV}}$. The solid black lines are the true coefficient functions.

Figure 24: Sensitivity of $\hat{h}_{\mathrm{CV}}$ to the choice of ν



For the simulation study in Section 5.3, this figure shows boxplots of $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$*, using the global bandwidth selector* $\hat{h}_{\mathrm{CV}}$ *with* $\nu^{(\%)} = 60$*,* 70*,* 80 *and* 90*.*

Figure 25: Sensitivity $\hat{k}_{\mathrm{CV}}$ to the choice of ν



For the simulation study in Section 5.3, this figure shows boxplots of $\Delta(\hat{\mathbf{g}}, \hat{\mathbf{\Phi}})$*, using the nearest neighbour bandwidth selector* $\hat{k}_{\mathrm{CV}}$ *with* $\nu^{(\%)} = 60$*,* 70*,* 80 *and* 90*.*

6 Introduction to FACE

In this chapter, we focus on the main subject of the thesis: estimation of high dimensional covariance matrices using our proposed dynamic structure. We use the abbreviation FACE (Factor model with an Adaptive-varying-coefficient-model structure Covariance matrix Estimator) to denote the estimator introduced in this section. This name was chosen because the estimator will 'face' the markets today based on what happened yesterday and adapt according to the dynamic structure.

6.1 Model specification

In this chapter we generalise the model structure explored in previous chapters by imposing a GARCH related structure on the idiosyncratic errors. This assumption is well suited to a financial setting, where it is common to witness continued periods of large volatility followed by continued periods of calm.

As before, assume that $\{(X_t, Y_t),\ t = 1, \cdots, n\}$ is a time series where Y_t denotes a vector of p_n response variables and X_t denotes a vector of q (observable) factors. We still assume that $p_n \longrightarrow \infty$ as $n \longrightarrow \infty$, and q is fixed, and that $\{X_t,\ t = 1, \cdots, n\}$ is a stationary Markov process. The proposed dynamic model structure is

$$Y_t = \mathbf{g}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta})X_t + \boldsymbol{\epsilon}_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \qquad (6.1)$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_q)^{\mathrm{T}}$ is an unknown direction vector, $\mathbf{g}(\cdot)$ is an unknown intercept vector, $\boldsymbol{\Phi}(\cdot)$ is an unknown factor loading matrix, and $\boldsymbol{\epsilon}_t = (\epsilon_{1,t}, \cdots, \epsilon_{p_n,t})^{\mathrm{T}}$ is a p_n-dimensional random error vector at time t. We assume $\{\boldsymbol{\epsilon}_t,\ t = 1, \cdots, n\}$ is independent of $\{X_t,\ t =$

$1, \cdots, n\}$, and we further assume that $E(\boldsymbol{\epsilon}_t | \{\boldsymbol{\epsilon}_l : l < t\}) = \mathbf{0}$ and that

$$\operatorname{cov}(\boldsymbol{\epsilon}_t | \{\boldsymbol{\epsilon}_l : l < t\}) = \boldsymbol{\Sigma}_{0,t} = \operatorname{diag}\{\sigma_{1,t}^2, \cdots, \sigma_{p_n,t}^2\}$$

where

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i} \epsilon_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j} \sigma_{k,t-j}^2, \quad t = 2, \cdots, n \qquad (6.2)$$

for each $k = 1, \cdots, p_n$ and for some integers m and s. Throughout the numerical studies of this thesis, we will simply choose $m = 1$ and $s = 1$.

Let $\mathcal{F}_t$ be the σ-algebra generated by $\{(X_l^{\mathrm{T}}, \boldsymbol{\epsilon}_l^{\mathrm{T}}) : l \leq t\}$. The main focus of this chapter is on the conditional covariance matrix $\operatorname{cov}(Y_t | \mathcal{F}_{t-1})$, for which an expression can be obtained by taking the conditional covariance of both sides of (6.1), yielding the crucial identity

$$\operatorname{cov}(Y_t | \mathcal{F}_{t-1}) = \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}} \boldsymbol{\beta}) \boldsymbol{\Sigma}_x(X_{t-1}) \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}} \boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{\Sigma}_{0,t} \qquad (6.3)$$

where $\boldsymbol{\Sigma}_x(X_{t-1}) \equiv \operatorname{cov}(X_t | X_{t-1})$. In a similar way, by taking conditional expectations, we have

$$E(Y_t | \mathcal{F}_{t-1}) = \mathbf{g}(X_{t-1}^{\mathrm{T}} \boldsymbol{\beta}) + \boldsymbol{\Phi}(X_{t-1}^{\mathrm{T}} \boldsymbol{\beta}) E(X_t | X_{t-1}). \qquad (6.4)$$

This will also be useful to us, since in Markowitz's formula for portfolio allocation, we need to estimate both the covariance matrix and the mean vector of excess asset returns.

6.2 Methodology

Estimating $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$ and $E(Y_t|\mathcal{F}_{t-1})$ using the following substitution estimators

$$\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1}) = \hat{\mathbf{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{\mathbf{\Sigma}}_x(X_{t-1})\hat{\mathbf{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})^{\mathrm{T}} + \hat{\mathbf{\Sigma}}_{0,t} \tag{6.5}$$

and

$$\hat{E}(Y_t|\mathcal{F}_{t-1}) = \hat{\mathbf{g}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \hat{\mathbf{\Phi}}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{E}(X_t|X_{t-1}) \tag{6.6}$$

can be broken down into the following steps:

(Step 1) Estimate $\boldsymbol{\beta}$ using either $\hat{\boldsymbol{\beta}}_{\mathrm{F}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$, or $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ as in Section 4.2.

(Step 2) Estimate $\mathbf{g}(\cdot)$ and $\mathbf{\Phi}(\cdot)$ using (4.4) and (4.5) using either the global bandwidth $\hat{h}_2$ or the nearest neighbour bandwidth $h_{\hat{k}}(z)$ as explained in Section 5.2.

(Step 3) Estimate $\mathbf{\Sigma}_{0,t}$ using methodology introduced in Section 6.2.1.

(Step 4) Estimate $\mathbf{\Sigma}_x(\cdot)$ using methodology introduced in Section 6.2.2.

Hence, in this section, we explore methodology for estimating $\mathbf{\Sigma}_x(\cdot)$ and $\mathbf{\Sigma}_{0,t}$. The recommended choice is to use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{L}}$ in Step 1 and a nearest neighbour bandwidth $h_{\hat{k}}(z)$ with $\hat{k}$ selected by cross validation in Step 2.

6.2.1 Estimation of $\mathbf{\Sigma}_{0,t}$

In this section, we introduce the methodology for estimating the idiosyncratic error covariance matrix $\mathbf{\Sigma}_{0,t}$. Denote the residuals by

$$r_{k,t} = \hat{\epsilon}_{k,t} = y_{k,t} - \hat{g}_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + X_t^{\mathrm{T}}\hat{\mathbf{a}}_k(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) \tag{6.7}$$

for each k, $k = 1, \cdots, p_n$. By substituting (6.7) into (6.2), we arrive at the following synthetic GARCH model

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i} r_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j} \sigma_{k,t-j}^2, \quad t = 2, \cdots, n \tag{6.8}$$

It can be shown that (6.8) has a re-parametrisation in the form of the following ARMA model

$$r_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{\max(m,s)} (\alpha_{k,i} + \gamma_{k,i}) r_{k,t-i}^2 + \eta_{k,t} - \sum_{j=1}^{s} \gamma_{k,j} \eta_{k,t-j}, \tag{6.9}$$

for each t, $t = 2, \cdots, n$, where $\eta_{k,t} = r_{k,t}^2 - \sigma_{k,t}^2$, $\gamma_{k,i} = 0$ when $i > s$, and $\alpha_{k,i} = 0$ when $i > m$.

It is now possible to use the estimation procedure for ARMA models in order to get estimates for $\alpha_{k,i}$ and $\gamma_{k,j}$. By substituting these estimates into (6.8) we can obtain an estimator $\hat{\sigma}_{k,t}^2$ of $\sigma_{k,t}^2$ and hence an estimator $\hat{\boldsymbol{\Sigma}}_{0,t}$ of $\boldsymbol{\Sigma}_{0,t}$.

For each k, $k = 1, \cdots, p_n$, let $\boldsymbol{\theta}_k = (\alpha_{k,0}, \cdots, \alpha_{k,m}, \gamma_{k,1}, \cdots, \gamma_{k,s})^{\mathrm{T}}$. We are going to use a quasi-maximum likelihood approach to estimate $\boldsymbol{\theta}_k$. We define the negative quasi log-likelihood function of $\boldsymbol{\theta}_k$ as

$$\mathcal{Q}_{k,n}(\boldsymbol{\theta}_k) = n^{-1} \sum_{t=2}^{n} \left\{ \frac{r_{k,t}^2}{\sigma_{k,t}^2(\boldsymbol{\theta}_k)} + \log \sigma_{k,t}^2(\boldsymbol{\theta}_k) \right\}$$

where $\sigma_{k,t}^2(\boldsymbol{\theta}_k)$ are recursively defined by (6.8) with initial values being either

$$r_{k,0}^2 = \cdots = r_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = \alpha_{k,0}$$

or

$$r_{k,0}^2 = \cdots = r_{k,1-m}^2 = \sigma_{k,0}^2 = \cdots = \sigma_{k,1-s}^2 = r_{k,0}^2.$$

By minimising $\mathcal{Q}_{k,n}(\boldsymbol{\theta}_k)$ with respect to $\boldsymbol{\theta}_k$ on a compact set

$$\boldsymbol{\Lambda} \subset (c, +\infty) \times (c, +\infty)^{m+s} \quad c > 0$$

we use the minimiser $\hat{\boldsymbol{\theta}}_k$ to estimate $\boldsymbol{\theta}_k$.

There are established methods for choosing m and s such as using AICC Brockwell and Davis (2002), although it is often practical to simply choose $m = 1$ and $s = 1$. To keep things straightforward, we simply assume $m = 1$ and $s = 1$ in the numerical studies in this thesis.

In order to implement the estimation procedure, and to simulate the data, one can take advantage of the `fGarch` R package by Wuertz and Chalabi (2013).

6.2.2 Estimation of $\boldsymbol{\Sigma}_x(\cdot)$

In this section, we introduce the methodology for estimating the conditional covariance matrix $\boldsymbol{\Sigma}_x(\cdot)$. A crucial remark is that, since q is fixed as $n \longrightarrow \infty$, the number of unknown parameters in $\boldsymbol{\Sigma}_x(\cdot)$ is significantly smaller than the number of unknown parameters in $\text{cov}(Y_t|\mathcal{F}_{t-1})$. This means the estimation for $\boldsymbol{\Sigma}_x(\cdot)$ does not suffer from the so-called 'curse of dimensionality'.

The proposed estimator is based on a local constant approximation. In order to estimate $E(X_t|X_{t-1} = \mathbf{u})$ and $E(X_t X_t^{\mathrm{T}}|X_{t-1} = \mathbf{u})$, for any given $\mathbf{u}$, we use the local constant estimators with

$$\widehat{E}(X_t|X_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n} X_t K_{h_3}(\|X_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n} K_{h_3}\left(\|X_{t-1} - \mathbf{u}\|\right)} \tag{6.10}$$

and

$$\widehat{E}(X_t X_t^{\mathrm{T}}|X_{t-1} = \mathbf{u}) = \frac{\sum_{t=2}^{n} X_t X_t^{\mathrm{T}} K_{h_3}(\|X_{t-1} - \mathbf{u}\|)}{\sum_{t=2}^{n} K_{h_3}\left(\|X_{t-1} - \mathbf{u}\|\right)} \tag{6.11}$$

where h_3 is a bandwidth. In both (6.10) and (6.11) we apply more weight to values of X_t when X_{t-1} is "close" in magnitude to $\mathbf{u}$. This gives us the following estimator of $\mathbf{\Sigma}_x(\mathbf{u})$

$$\begin{aligned}\hat{\mathbf{\Sigma}}_x(\mathbf{u}) =& \widehat{E}(X_tX_t^{\mathrm{T}}|X_{t-1}=\mathbf{u}) - \widehat{E}(X_t|X_{t-1}=\mathbf{u})\{\widehat{E}(X_t|X_{t-1}=\mathbf{u})\}^{\mathrm{T}} \\ =& \{\mathrm{tr}(\mathcal{W})\}^{-2}\mathbf{X}^{\mathrm{T}}\{\mathrm{tr}(\mathcal{W})\mathcal{W} - \mathcal{W}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathcal{W}\}\mathbf{X} \qquad (6.12)\end{aligned}$$

where

$$\mathbf{X} = (X_2, \cdots, X_n)^{\mathrm{T}}, \quad \mathcal{W} = \mathrm{diag}(K_{h_3}(\|X_1 - \mathbf{u}\|), \cdots, K_{h_3}(\|X_{n-1} - \mathbf{u}\|)).$$

Indeed, it can be seen that $\hat{\mathbf{\Sigma}}_x(\mathbf{u})$ is related to the sample covariance matrix, and they are identical to one another if $h_3 \longrightarrow \infty$.

We can choose h_3 using cross validation in a similar way to the previous chapter. First, fix $\nu < n$. We start by defining the cross validation statistic at time t for bandwidth h_3 by

$$\mathrm{CV}(h_3) = \sum_{t=n-\nu}^{n} \|X_tX_t^{\mathrm{T}} - \hat{E}^{(t-1)}(X_tX_t^{\mathrm{T}}|X_{t-1})\| + \|X_t - \hat{E}^{(t-1)}(X_t|X_{t-1})\|$$

where $\hat{E}^{(t-1)}(X_t|X_{t-1})$ and $\hat{E}^{(t-1)}(X_tX_t^{\mathrm{T}}|X_{t-1})$ are the respective estimates of $\widehat{E}(X_t|X_{t-1})$ and $\widehat{E}(X_tX_t^{\mathrm{T}}|X_{t-1})$ based on $\{X_l^{\mathrm{T}}, l = 1, \cdots, t-1\}$, and where ν is a look-back integer such that $\nu < n-1$. Hence, denoting the h_3 that minimises $\mathrm{CV}(h_3)$ by $\hat{h}_3$, we use the global bandwidth $\hat{h}_3$ in the local constant estimation of (6.10) and (6.11).

6.3 Portfolio allocation

In this section, we will briefly describe the construction of an optimum portfolio allocation based on the proposed dynamic structure and the associated estimation procedure. Our proposed portfolio allocation builds on the mean-variance portfolio by Markowitz (1952, 1968).

The allocation vector $\mathbf{w}$ of p_n risky assets, to be held between times $t-1$ and t, is defined as the solution to

$$\begin{aligned}&\min_{\mathbf{w}}\ \mathbf{w}^{\mathrm{T}}\mathrm{cov}(Y_t|\mathcal{F}_{t-1})\mathbf{w}\\ &\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1}_{p_n}=1 \quad \text{and} \quad \mathbf{w}^{\mathrm{T}}E(Y_t|\mathcal{F}_{t-1})=\delta\end{aligned}$$

where δ is the target return imposed on the portfolio. The solution $\hat{\mathbf{w}}$ is given by

$$\hat{\mathbf{w}}=\frac{c_3-c_2\delta}{c_1c_3-c_2^2}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n}+\frac{c_1\delta-c_2}{c_1c_3-c_2^2}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\hat{E}(Y_t|\mathcal{F}_{t-1})$$

where

$$c_1=\mathbf{1}_{p_n}^{\mathrm{T}}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\mathbf{1}_{p_n}, \quad c_2=\mathbf{1}_{p_n}^{\mathrm{T}}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\hat{E}(Y_t|\mathcal{F}_{t-1})$$

$$c_3=\hat{E}(Y_t|\mathcal{F}_{t-1})^{\mathrm{T}}\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1})^{-1}\hat{E}(Y_t|\mathcal{F}_{t-1}).$$

We remark that one can impose additional constraints to the above optimisation problem, and the solution can be solved for numerically. An example relating to gross exposure constraints will be presented in Section 7.6. The key point is that, regardless of the additional constraints, the estimated portfolio always requires a good estimator of the covariance matrix.

6.4 Simulation study

We now examine the performance of FACE using a simulated example. In Section 6.4.1 we explore the sensitivity of FACE to the choice of $\hat{\boldsymbol{\beta}}$ and $\hat{h}_2$, and make comparisons with the conclusions from previous chapters. In Section 6.4.2 we compare FACE to some other commonly used estimators such as the sample covariance matrix estimator and the estimator proposed by Fan et al. (2008a).

6.4.1 Sensitivity to the choice of $\hat{\boldsymbol{\beta}}$ and $\hat{h}_2$

We start by focusing on how the choice of estimators for $\boldsymbol{\beta}$ and h_2 affects the quality of estimators $\widehat{\text{cov}}(Y_t|\mathcal{F}_{t-1})$ and $\hat{E}(Y_t|\mathcal{F}_{t-1})$. We generate 1000 datasets from model (6.1), together with (6.2), each with a sample size of $n = 1000$ and $p_n = 50$. We set $q = 4$, $m = 1$, and $s = 1$. For $k = 1, \cdots, p_n$, we set $\alpha_{0,k} = 0.5$, $\alpha_{1,k} = 0.1$, and $\beta_{1,k} = 0.1$. We set $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ equal to the same values as those from Section 4.3. For $t = 1, \cdots, n+1$, we generate X_t independently from a uniform distribution on $[-1, 1]^q$, Z_t from a p_n-variate standard normal distribution, and $\boldsymbol{\epsilon}_t$ through $\boldsymbol{\epsilon}_t = \boldsymbol{\Sigma}_{0,t}^{1/2} Z_t$. Once X_t and $\boldsymbol{\epsilon}_t$ have been generated, Y_t can be generated through (6.1) for $t = 1, \cdots, n+1$.

We will initially pretend that $(X_{n+1}^{\mathrm{T}}, Y_{n+1}^{\mathrm{T}})$ is unknown to us, and this will not be used in the estimation of $\text{cov}(Y_t|\mathcal{F}_{t-1})$ and $E(Y_t|\mathcal{F}_{t-1})$. The purpose of generating an additional data point $(X_{n+1}^{\mathrm{T}}, Y_{n+1}^{\mathrm{T}})$ is to enable us to calculate the one-period simple return

$$R(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}} Y_{n+1}$$

of a portfolio allocation $\hat{\mathbf{w}}$ formed at time n, based on data $(X_t^{\mathrm{T}}, Y_t^{\mathrm{T}})$, $t = 1, \cdots, n$, and held until time $n + 1$. We use the Sharpe ratio

$$\text{SR}(\hat{\mathbf{w}}) = \frac{E(R(\hat{\mathbf{w}}))}{\text{SD}(R(\hat{\mathbf{w}}))}$$

to evaluate the performance of a portfolio allocation $\hat{\mathbf{w}}$, where $\text{SD}(R(\hat{\mathbf{w}}))$ is the standard deviation of $R(\hat{\mathbf{w}})$. For simplicity we assume a zero risk-free rate and use a target return of $\delta = 1\%$, although similar conclusions can be made for other positive values of δ.

We use the metrics

$$\Delta(\hat{M}, M) = \frac{\|\hat{M} - M\|_{\mathrm{F}}}{\|M\|_{\mathrm{F}}}, \quad \epsilon(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \frac{1}{p_n}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_{\mathrm{F}}$$

to evaluate the performance of the estimator $\hat{M}$ of matrix M and the estimator $\hat{\boldsymbol{\mu}}$ of a p_n-dimensional column vector $\boldsymbol{\mu}$, where

$$\|M\|_{\mathrm{F}} = \{\mathrm{tr}(MM^{\mathrm{T}})\}^{1/2}$$

is the Frobenius norm. A crucial remark is that both $\Delta(\hat{M}, M)$ and $\epsilon(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu})$ can only be used for simulated data, but not with real data since the true parameters are unknown in practice. Using $\mathrm{SR}(\hat{\mathbf{w}})$ has the advantage that it measures the quality of estimation for real data too and, more importantly, is relevant to an investor from a financial point of view. The Sharpe ratio can be viewed as a reward-to-risk ratio, and so larger values of $\mathrm{SR}(\hat{\mathbf{w}})$ are desirable.

For each generated dataset, we estimate $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$ and $E(Y_t|\mathcal{F}_{t-1})$ using different combinations of

$$\hat{\boldsymbol{\beta}} \in \{\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}},\ \hat{\boldsymbol{\beta}}_{\mathrm{L}},\ \hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}\}$$

and

$$\hat{h}_2 \in \{\hat{h}_{\mathrm{AIC}},\ \hat{h}_{\mathrm{BIC}},\ \hat{h}_{\mathrm{CV}},\ h_{\hat{k}_{\mathrm{AIC}}}(\cdot),\ h_{\hat{k}_{\mathrm{BIC}}}(\cdot),\ h_{\hat{k}_{\mathrm{CV}}}(\cdot)\}.$$

For each combination of $\hat{\boldsymbol{\beta}}$ and $\hat{h}_2$, we compute means and standard deviations of the following quantities

$$\Delta(\hat{\boldsymbol{\beta}}), \quad \Delta(\hat{\boldsymbol{\Phi}}(X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}})), \quad \Delta(\hat{\boldsymbol{\Sigma}}_x(X_n)), \quad \Delta(\hat{\boldsymbol{\Sigma}}_{0,n+1}), \quad R(\hat{\mathbf{w}}),$$

$$\Delta(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}, \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1}), \quad \epsilon(\hat{E}(Y_{n+1}|\mathcal{F}_n), E(Y_{n+1}|\mathcal{F}_n))$$

over the 1000 replications.

For the estimator of $\mathbf{\Sigma}_x(\cdot)$, we note that there is no dependence on $\hat{h}_2$ and $\boldsymbol{\beta}$, so in all cases we have

$$E(\Delta(\hat{\mathbf{\Sigma}}_x(X_n))) = 0.082, \quad \mathrm{SD}(\Delta(\hat{\mathbf{\Sigma}}_x(X_n))) = 0.0413.$$

As expected, we do not suffer from the 'curse of dimensionality' because $\mathbf{\Sigma}_x(\cdot)$ is only a $q \times q$ matrix.

A comparison of the performance of the substituted quantities $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{\Phi}}(X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}})$, and $\hat{\mathbf{\Sigma}}_{0,n+1}$, as well as $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}$ and $\hat{E}(Y_{n+1}|\mathcal{F}_n)$, can be found in Tables 6.1, 6.2 and 6.3. In general, these findings agree with the conclusions from Chapters 3, 4 and 5. In particular it is possible to estimate $\boldsymbol{\beta}$ within a reasonable degree of accuracy using either $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ or $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$, and both of these estimators outperform $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$. We also see that the Sharpe ratios using $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are significantly higher than $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ too. Additionally we also notice that bandwidths selected using $\hat{k}_{\mathrm{AIC}}$ and $\hat{k}_{\mathrm{BIC}}$ still perform poorly, which was noted in Chapter 5. The recommended method for selecting a bandwidth, using $\hat{k}_{\mathrm{CV}}$ works well as before, and we see that $\hat{h}_{\mathrm{CV}}$ has improved slightly in this simulation study compared to Section 5.3.

It is possible to gain more insight into these estimators from the boxplots in Figure 26. As previously mentioned, $\hat{h}_{\mathrm{CV}}$ can perform quite well most of the time, especially when the effects of the boundary are not too severe. However we notice from the boxplot that there are several outliers corresponding to large approximation errors resulting most likely from errors at the boundary caused by remote data points. On the other hand, $\hat{k}_{\mathrm{CV}}$ is more stable and does not have such large outliers. This is a key reason why we propose using $\hat{k}_{\mathrm{CV}}$ to select the bandwidth.

Similar conclusions can also be made from the boxplots of portfolio returns in Figure 27. For example, we see that the realised returns are more concentrated around the target rate $\delta = 1\%$ when $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are used, and more dispersed if $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$ is used.

In summary, we can confirm the key conclusions made in previous chapters also hold in the context for covariance matrix estimation and portfolio allocation. In particular, for the estimation procedure in Section 6.2, we recommend using $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{L}}$ in Step 1 and a nearest neighbour bandwidth $h_{\hat{k}}(z)$ with $\hat{k}$ selected by cross validation in Step 2. These choices will be used in the Section 6.4.2, as well as in the real data analysis in Chapter 7.

Table 6.1: Performance of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Phi}}(X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}})$

		$E[D_1]$	$\mathrm{SD}[D_1]$	$E[D_2]$	$\mathrm{SD}[D_2]$
$\hat{h}_{\mathrm{AIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.41	0.19
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.29	0.09
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.29	0.09
$h_{\hat{k}_{\mathrm{AIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.49	0.27
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.45	0.28
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.45	0.28
$\hat{h}_{\mathrm{BIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.42	0.19
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.31	0.10
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.31	0.10
$h_{\hat{k}_{\mathrm{BIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.49	0.26
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.46	0.28
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.46	0.28
$\hat{h}_{\mathrm{CV}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.35	0.23
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.16	0.10
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.16	0.10
$h_{\hat{k}_{\mathrm{CV}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.59	0.45	0.35	0.23
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.01	0.01	0.15	0.04
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.01	0.01	0.15	0.04

For the simulation study in Section 6.4.1, this table shows the mean and standard deviation of relative error metrics $D_1 = \Delta(\hat{\boldsymbol{\beta}})$ and $D_2 = \Delta(\hat{\boldsymbol{\Phi}}(X_n^{\mathrm{T}}\hat{\boldsymbol{\beta}}))$. The first column shows the choice of data driven bandwidth $\hat{h}_2$ and the second column shows the choice of estimator for $\boldsymbol{\beta}$.

Table 6.2: Performance of $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}$

		$E[D_3]$	$\mathrm{SD}[D_3]$	$E[D_4]$	$\mathrm{SD}[D_4]$	$E[D_5]$	$\mathrm{SD}[D_5]$
$\hat{h}_{\mathrm{AIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.49	0.33	0.32	0.13	0.18	0.04
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.18	0.03	0.16	0.02	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.18	0.03	0.16	0.02	0.17	0.04
$h_{\hat{k}_{\mathrm{AIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.88	0.21	0.48	0.06	0.18	0.05
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.79	0.13	0.45	0.04	0.18	0.05
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.79	0.13	0.45	0.04	0.18	0.05
$\hat{h}_{\mathrm{BIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.51	0.33	0.33	0.13	0.18	0.04
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.20	0.04	0.18	0.02	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.20	0.04	0.18	0.02	0.17	0.04
$h_{\hat{k}_{\mathrm{BIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.89	0.21	0.48	0.06	0.18	0.05
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.80	0.13	0.46	0.04	0.18	0.05
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.80	0.13	0.46	0.04	0.18	0.05
$\hat{h}_{\mathrm{CV}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.40	0.34	0.27	0.15	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.10	0.02	0.11	0.03	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.10	0.02	0.11	0.03	0.17	0.04
$h_{\hat{k}_{\mathrm{CV}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$	0.38	0.34	0.26	0.15	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\mathrm{L}}$	0.10	0.02	0.11	0.02	0.17	0.04
	$\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$	0.10	0.02	0.11	0.02	0.17	0.04

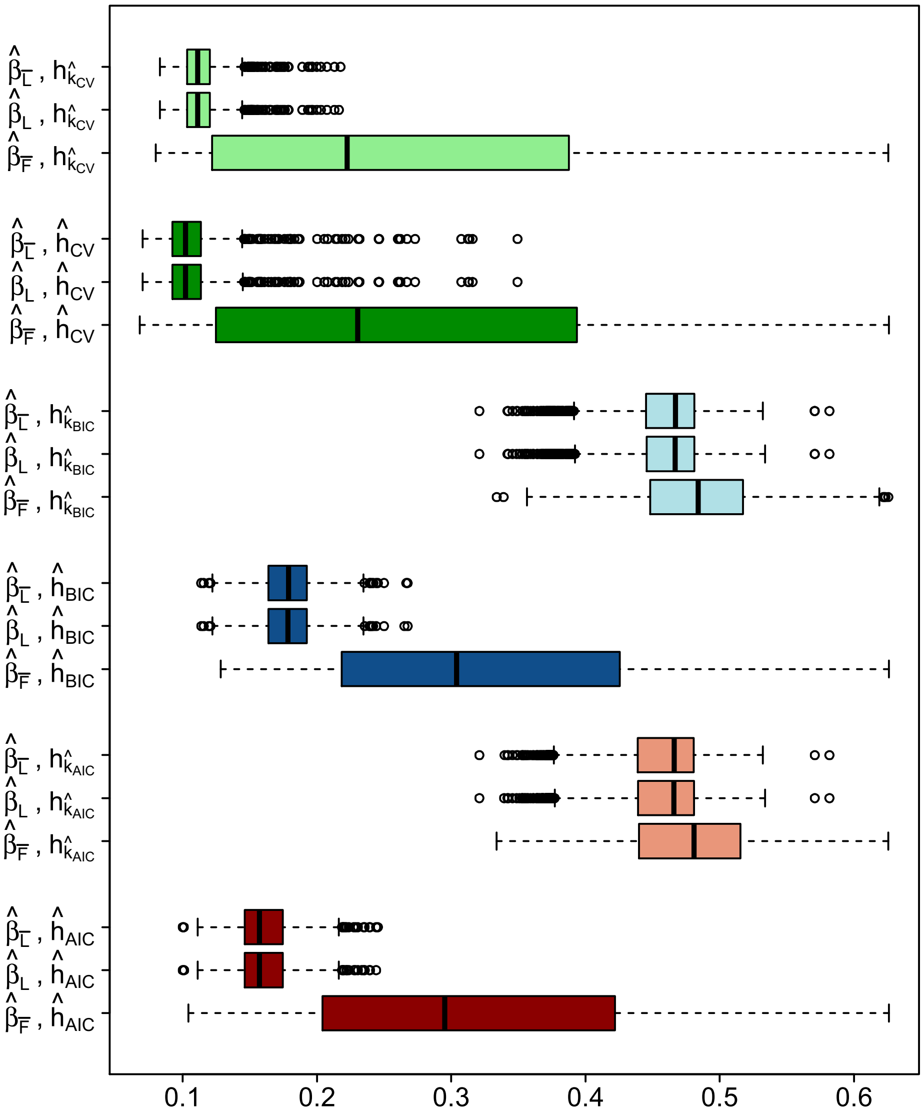
For the simulation study in Section 6.4.1, this table shows the mean and standard deviation of relative error metrics $D_3 = \Delta(\hat{\boldsymbol{\Sigma}}_{0,n+1})$, $D_4 = \Delta(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}, \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1})$ and $D_5 = \epsilon(\hat{E}(Y_{n+1}|\mathcal{F}_n), E(Y_{n+1}|\mathcal{F}_n))$. The first column shows the choice of data driven bandwidth $\hat{h}_2$ and the second column shows the choice of estimator for $\boldsymbol{\beta}$.

Table 6.3: Performance of $\hat{\mathbf{w}}$

		$E[R(\hat{\mathbf{w}}))]$	$\mathrm{SD}[R(\hat{\mathbf{w}}))]$	$SR[R(\hat{\mathbf{w}})]$
$\hat{h}_{\text{AIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	1.10	0.65	1.7
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	1.13	0.47	2.4
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	1.13	0.47	2.4
$h_{\hat{k}_{\text{AIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	1.02	0.81	1.3
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	1.01	0.77	1.3
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	1.01	0.77	1.3
$\hat{h}_{\text{BIC}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	1.11	0.66	1.7
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	1.14	0.49	2.4
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	1.15	0.49	2.4
$h_{\hat{k}_{\text{BIC}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	1.02	0.81	1.3
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	1.01	0.78	1.3
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	1.01	0.78	1.3
$\hat{h}_{\text{CV}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	1.00	0.62	1.6
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	1.01	0.41	2.5
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	1.01	0.41	2.5
$h_{\hat{k}_{\text{CV}}}$	$\hat{\boldsymbol{\beta}}_{\bar{\text{F}}}$	0.98	0.62	1.6
	$\hat{\boldsymbol{\beta}}_{\text{L}}$	0.99	0.40	2.5
	$\hat{\boldsymbol{\beta}}_{\bar{\text{L}}}$	0.99	0.40	2.5

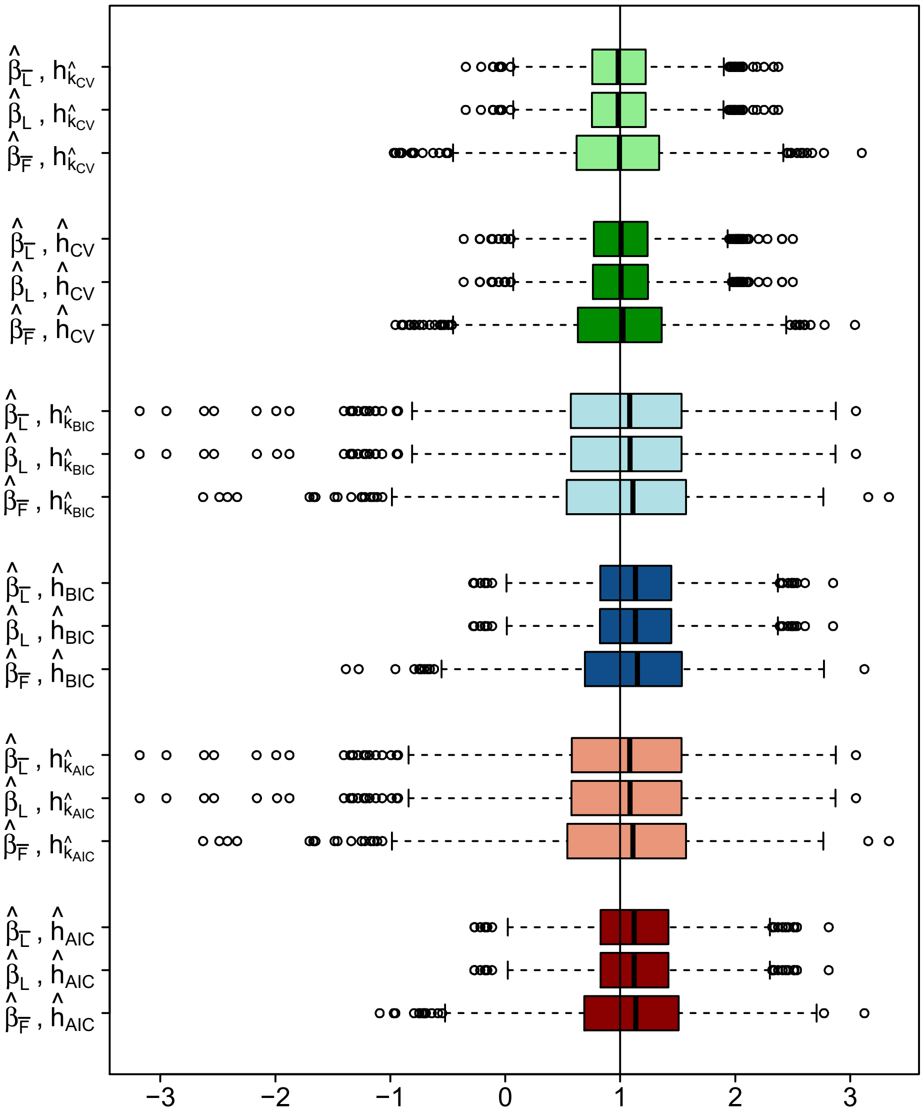
For the simulation study in Section 6.4.1, this table shows the mean and standard deviation of the portfolio returns $R(\hat{\mathbf{w}})$, along with the Sharpe ratio $SR[R(\hat{\mathbf{w}})]$. The first column shows the choice of data driven bandwidth $\hat{h}_2$ and the second column shows the choice of estimator for $\boldsymbol{\beta}$.

Figure 26: Estimation errors of $\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}$



For the simulation study in Section 6.4.1, this figure shows boxplots of the error metric $\Delta(\widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)^{-1}, \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1})$ over the 1000 replications, grouped according to the different combinations of $\hat{h}_2$ and $\hat{\boldsymbol{\beta}}$.

Figure 27: One-period simple returns



For the simulation study in Section 6.4.1, this figure shows boxplots of the one-period returns $R(\hat{\mathbf{w}})$ over the 1000 replications, grouped according to the different combinations of $\hat{h}_2$ and $\hat{\boldsymbol{\beta}}$. Returns are reported as percentages.

6.4.2 Alternative estimators

In this section we compare FACE with some other commonly used estimators using simulated data. We generate 1000 datasets from model (6.1), together with (6.2), using the same simulation settings as Section 6.4.1. We repeat this using the following combinations of n and p_n: $\{n = 1000,\ p_n = 50\}$, $\{n = 1000,\ p_n = 100\}$, $\{n = 2000,\ p_n = 50\}$, and $\{n = 2000,\ p_n = 100\}$.

We shall use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and a nearest neighbour bandwidth $h_{\hat{k}}(z)$ with $\hat{k}$ selected by cross validation to estimate $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$, and denote our proposed estimators, based on the dynamic structure, by $\hat{\boldsymbol{\Sigma}}_{\mathrm{FACE}} = \widehat{\mathrm{cov}}(Y_{n+1}|\mathcal{F}_n)$ and $\hat{\mathbf{w}}_{\mathrm{FACE}} = \hat{\mathbf{w}}$ as described in Section 6.2 and Section 6.3 respectively.

An alternative approach to estimating $\mathrm{cov}(Y_{n+1}|\mathcal{F}_n)$ is to simply ignore the dynamic structure and use the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{SAM}} = (n-1)^{-1}\mathbf{Y}\mathbf{Y}^{\mathrm{T}} - \{n(n-1)\}^{-1}\mathbf{Y}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}$$

where $\mathbf{Y} = (Y_1, \cdots, Y_n)$. We can use the sample mean and $\hat{\boldsymbol{\Sigma}}_{\mathrm{SAM}}$ in Markowitz's formula (see Section 6.3) to form an estimated portfolio allocation which we denote by $\hat{\mathbf{w}}_{\mathrm{SAM}}$.

In a similar way, one can ignore the dynamic structure, but use a factor model whereby the coefficients are assumed constant. Still using the $q = 4$ factors, we explore the covariance matrix estimator based the linear factor model explored by Fan et al. (2008a), which was summarized in the literature review in Section 2.3. We denoted this estimator by $\hat{\boldsymbol{\Sigma}}_{\mathrm{FAN}}$, and its corresponding estimated portfolio allocation by $\hat{\mathbf{w}}_{\mathrm{FAN}}$.

In Table 6.4, we make a comparison of $\hat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{FACE}}$, $\hat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{SAM}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{FAN}}$ in terms of $\Delta(\hat{\boldsymbol{\Sigma}}^{-1}, \mathrm{cov}(Y_{n+1}|\mathcal{F}_n)^{-1})$ as well as their realised portfolio returns and Sharpe ratios. It is easy to see that $\hat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{FACE}}$ greatly outperforms $\hat{\mathbf{\Sigma}}_{\text{SAM}}^{-1}$ and $\hat{\mathbf{\Sigma}}_{\text{FAN}}^{-1}$. A more graphical and easier way of comparing the three estimators is given in Figure 28, which shows how the relative errors are distributed across the 1000 replications for each combination of n and p_n. The crucial point here, is that if one ignores the dynamic structure, and uses $\hat{\mathbf{\Sigma}}_{\text{SAM}}^{-1}$ and $\hat{\mathbf{\Sigma}}_{\text{FAN}}^{-1}$, then the approximation errors are strikingly large. However, we see that by taking into account the dynamic structure, by using $\hat{\mathbf{\Sigma}}_{\text{FACE}}^{-1}$, we can achieve a much better quality of estimation.

We also see that when p_n increases, but n stays fixed, that $\hat{\mathbf{\Sigma}}_{\text{SAM}}^{-1}$ suffers from increased errors due to an increase in number of parameters to estimate. Indeed, this is a well known fact about the sample covariance matrix which this simulation study has verified. However, the vital point is that increasing p_n does not worsen the estimation quality of $\hat{\mathbf{\Sigma}}_{\text{FACE}}^{-1}$, since the proposed dynamic structure successfully reduces the dimension.

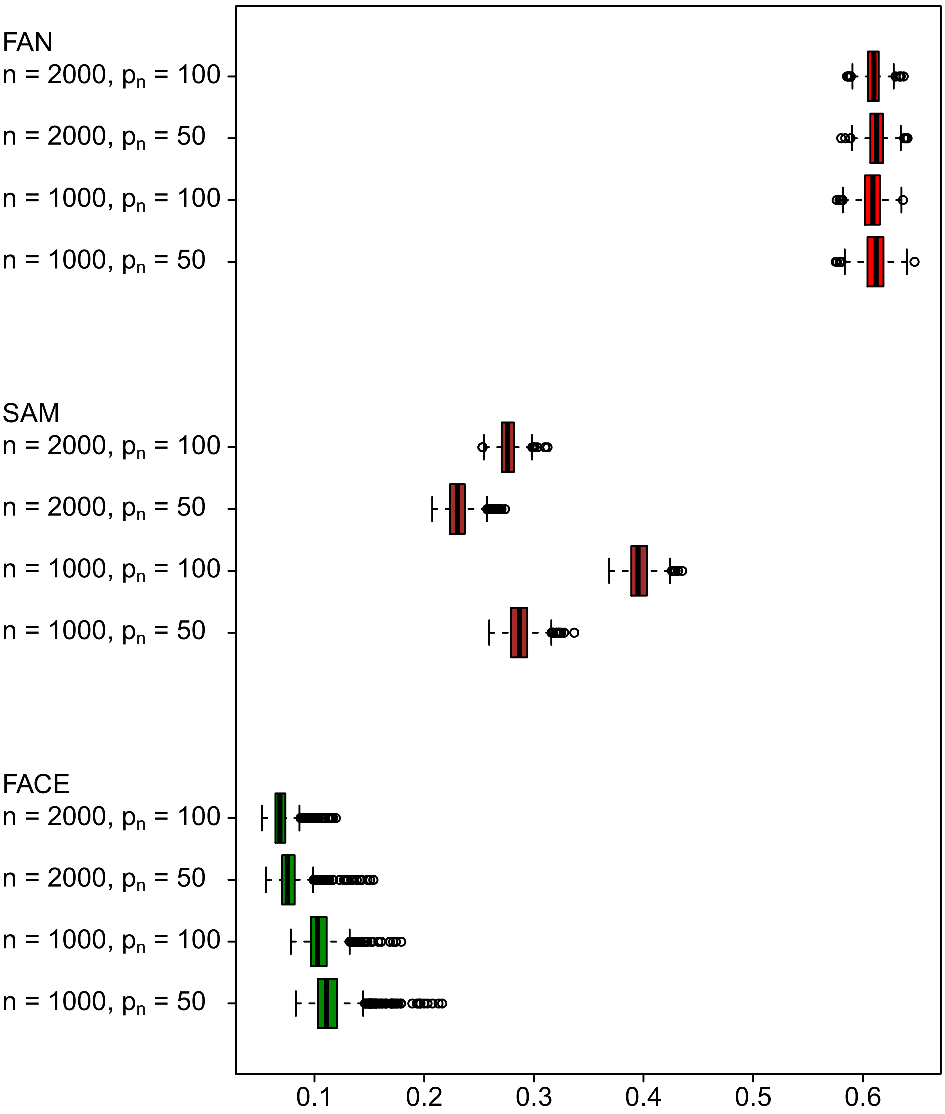
These observations can also be seen by the fact that the Sharpe ratios $\text{SR}(\hat{\mathbf{w}}_{\text{FACE}})$ are notably higher than $\text{SR}(\hat{\mathbf{w}}_{\text{SAM}})$ and $\text{SR}(\hat{\mathbf{w}}_{\text{FAN}})$ for all combinations of n and p_n. Finally, we also remark that the portfolio returns resulted from $\hat{\mathbf{w}}_{\text{FACE}}$ are relatively closely distributed around the target rate $\delta = 1\%$ whereas $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ have returns which are more dispersed.

Table 6.4: Comparison of $\hat{\mathbf{\Sigma}}_{\text{FACE}}$, $\hat{\mathbf{\Sigma}}_{\text{SAM}}$ and $\hat{\mathbf{\Sigma}}_{\text{FAN}}$

		$n = 1000$, $p_n = 50$	$n = 1000$, $p_n = 100$	$n = 2000$, $p_n = 50$	$n = 2000$, $p_n = 100$
$E[D]$	FACE	0.11	0.11	0.08	0.07
	SAM	0.29	0.40	0.23	0.28
	FAN	0.61	0.61	0.61	0.61
$\text{SD}[D]$	FACE	0.02	0.01	0.01	0.01
	SAM	0.01	0.01	0.01	0.01
	FAN	0.01	0.01	0.01	0.01
$E[R(\hat{\mathbf{w}})]$	FACE	0.99	1.01	1.03	1.03
	SAM	0.96	0.96	1.02	1.02
	FAN	0.96	0.96	1.02	1.02
$\text{SD}[R(\hat{\mathbf{w}})]$	FACE	0.40	0.28	0.39	0.27
	SAM	1.02	1.03	1.03	1.02
	FAN	0.99	0.97	1.02	1.00
$\text{SR}(\hat{\mathbf{w}})$	FACE	2.5	3.6	2.6	3.8
	SAM	0.9	0.9	1.0	1.0
	FAN	1.0	1.0	1.0	1.0

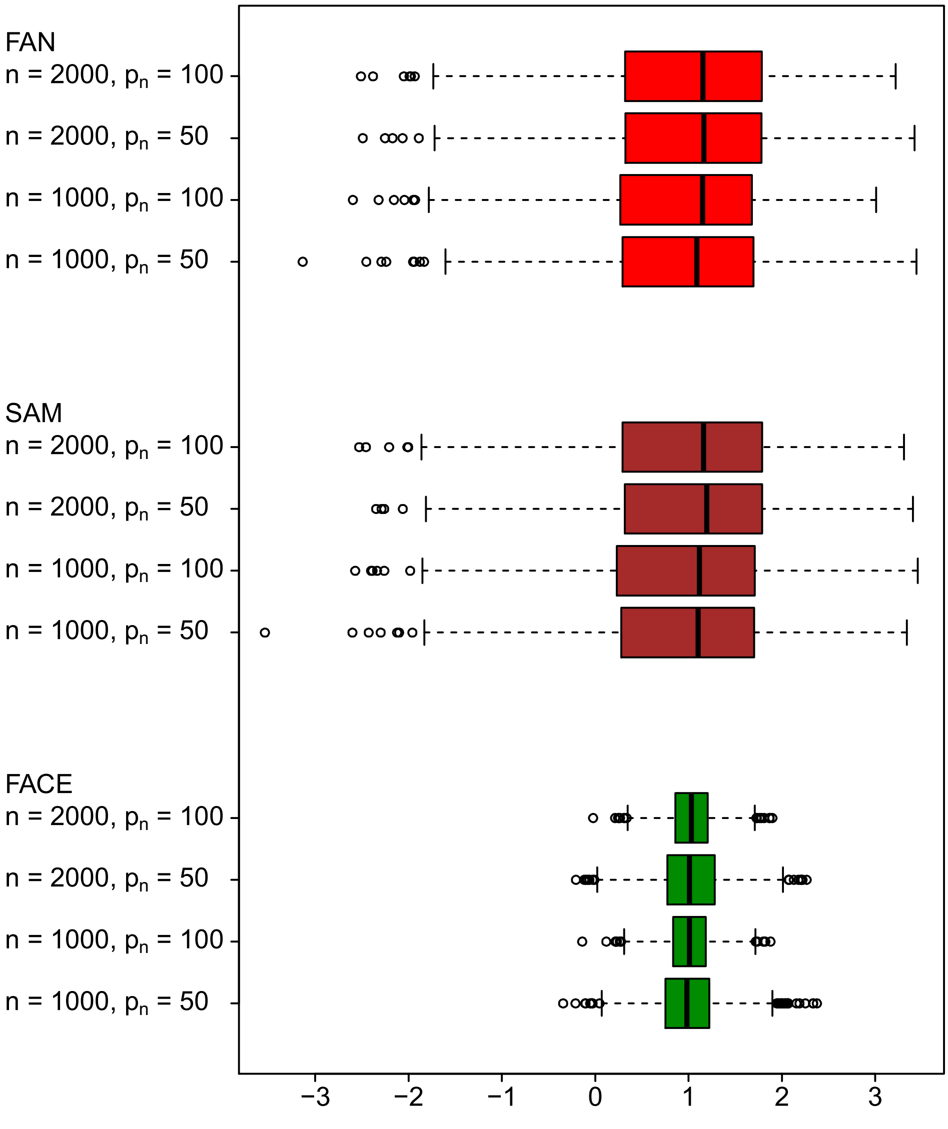
For the simulation study in Section 6.4.2, this table shows the mean and standard deviation of $D = \Delta(\hat{\mathbf{\Sigma}}^{-1}, \text{cov}(Y_{n+1}|\mathcal{F}_n)^{-1})$ for $\hat{\mathbf{\Sigma}}^{-1} = \hat{\mathbf{\Sigma}}^{-1}_{\text{FACE}}$, $\hat{\mathbf{\Sigma}}^{-1}_{\text{SAM}}$, and $\hat{\mathbf{\Sigma}}^{-1}_{\text{FAN}}$, along with the portfolio returns $R(\hat{\mathbf{w}})$, for $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$, and $\hat{\mathbf{w}}_{\text{FAN}}$. The results are grouped according to different combinations of n and p_n, and Sharpe ratios are also recorded in the final three rows.

Figure 28: Comparison of $\hat{\Sigma}^{-1}_{\text{FACE}}$, $\hat{\Sigma}^{-1}_{\text{SAM}}$ and $\hat{\Sigma}^{-1}_{\text{FAN}}$



For the simulation study in Section 6.4.2, this figure shows boxplots over the 1000 replications of relative error metrics $\Delta(\hat{\Sigma}^{-1}, \text{cov}(Y_{n+1}|\mathcal{F}_n)^{-1})$ *for* $\hat{\Sigma}^{-1} = \hat{\Sigma}^{-1}_{\text{FACE}}$, $\hat{\Sigma}^{-1}_{\text{SAM}}$, *and* $\hat{\Sigma}^{-1}_{\text{FAN}}$. *The results are grouped according to different combinations of* n *and* p_n.

Figure 29: Comparison of $\hat{\mathbf{w}}_{\mathrm{FACE}}$, $\hat{\mathbf{w}}_{\mathrm{SAM}}$ and $\hat{\mathbf{w}}_{\mathrm{FAN}}$



For the simulation study in Section 6.4.2, this figure shows boxplots over the 1000 replications of one-period simple returns $R(\hat{\mathbf{w}})$ for $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{FACE}}$, $\hat{\mathbf{w}}_{\mathrm{SAM}}$, and $\hat{\mathbf{w}}_{\mathrm{FAN}}$. The results are grouped according to different combinations of n and p_n.

7 Real Data Analysis

In this chapter we are going to apply the dynamic structure for covariance matrices to four real datasets. We compare $\hat{\mathbf{w}}_{\text{FACE}}$ with the allocation based on the sample covariance matrix (denoted by $\hat{\mathbf{w}}_{\text{SAM}}$), and the allocation proposed by Fan, Fan and Lv (2008) (denoted by $\hat{\mathbf{w}}_{\text{FAN}}$), which were described in Section 6.4.2.

7.1 Description of datasets

In this section, we provide a brief description of the real datasets which we use to compare the proposed methodology. All data can be freely downloaded from Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` and was accessed on 2nd April 2015.

The observable factors $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$ are taken to be the (daily) market, size and value factors respectively from the Fama-French three-factor model, as described in Table 7.1. The response variable Y_t is chosen to be (daily) simple returns from one of four datasets, which we shall separately analyse, minus the risk-free rate. The risk-free rate is taken to be the one-month Treasury bill rate and is also included on Kenneth French's website. The labeling along with a brief description can be found in Table 7.2. It is important to test the proposed methodology, $\hat{\mathbf{w}}_{\text{FACE}}$, on multiple datasets and across such a wide time period because in some periods, any trading strategy may perform well or poorly simply due to chance alone. By using twenty years (approximately 5000 trading days) worth of data, and independently analysing four datasets, we hope to improve the reliability of our analysis.

There are various advantages of using the (excess) portfolio returns for $y_{k,t}$ as opposed to using individual stocks: we avoid having to merge different sources of data; we avoid survivorship bias (where we only picked companies that did not go bankrupt); and we attempt to avoid company specific risk. A further benefit is that the data is free and presented in a spreadsheet format. These results can be independently reproduced using the C++, R and Bash source code found on `www.johnleighbox.co.uk`.

To have a better idea about what the data is like, we plot the observations from 3rd January 1995 to 31st December 2014 of the three factors and the risk-free rate in Figure 30, and the first four components of Y_t (from Dataset 1) in Figure 31 corresponding to the industrial sectors: Agriculture, Food Products, Candy & Soda, and Beer & Liquor. The plots show clearly that there are periods of large volatility around the 2008-2009 financial crisis. We will see $\hat{\mathbf{w}}_{\mathrm{FACE}}$ performs reasonably well even during that period, whilst $\hat{\mathbf{w}}_{\mathrm{SAM}}$ and $\hat{\mathbf{w}}_{\mathrm{FAN}}$ do not.

We initially focus on Dataset 1 (49 Industry Portfolios) where each $y_{k,t}$ can be easily interpreted as a portfolio consisting of company shares from a given industry. To get a better understanding of how certain industries react to periods of financial instability, we plot the cumulative returns of some industrial portfolios in Figure 32. For example, during the collapse of the dot-com bubble in 1999-2001, we see that technology related industry portfolios, Softw (Computer Software), Chips (Electronic Equipment) and LabEq (Measuring and Control Equipment) all suffered significant losses. Also, during the 2008-2009 financial crisis, financial related industry portfolios, such as Banks (Banking), Insur (Insurance) and Fin (Trading), experienced large drawdowns. However, we see that some industry portfolios were only marginally affected, if at all, by both of these periods of financial volatility, such as: Fun (Entertainment), Beer (Beer & Liquor), Toys (Recreation) and Smoke (Tobacco Products). This preliminary data analysis shows strong motivation for dynamically allocating a portfolio using the proposed methodology.

Table 7.1: Description of datasets for X_t

j	Name	Description
1	Market factor	Return on the market minus the risk-free rate
2	Size factor	Excess returns of small caps over big caps
3	Value factor	Excess returns of value stocks over growth stocks

This table gives name of factor j, corresponding to $x_{j,t}$, and a brief description.

Table 7.2: Description of datasets for Y_t

j	Name	Brief description	p_n
1	49 Industry Portfolios	Each NYSE, AMEX, and NASDAQ stock is assigned to one of 49 industry portfolios based on its four-digit Compustat SIC code.	49
2	100 Portfolios Formed on Size and Book-to- Market	Intersection of 10 portfolios formed on size (market equity, ME) and 10 portfolios formed on the ratio of book equity to market equity (BE/ME).	97
3	100 Portfolios Formed on Size and Investment	Intersection of 10 portfolios formed on size (market equity, ME) and 10 portfolios formed on investment (Inv).	99
4	100 Portfolios Formed on Size and Operating Profitability	Intersection of 10 portfolios formed on size (market equity, ME) and 10 portfolios formed on profitability (OP).	99

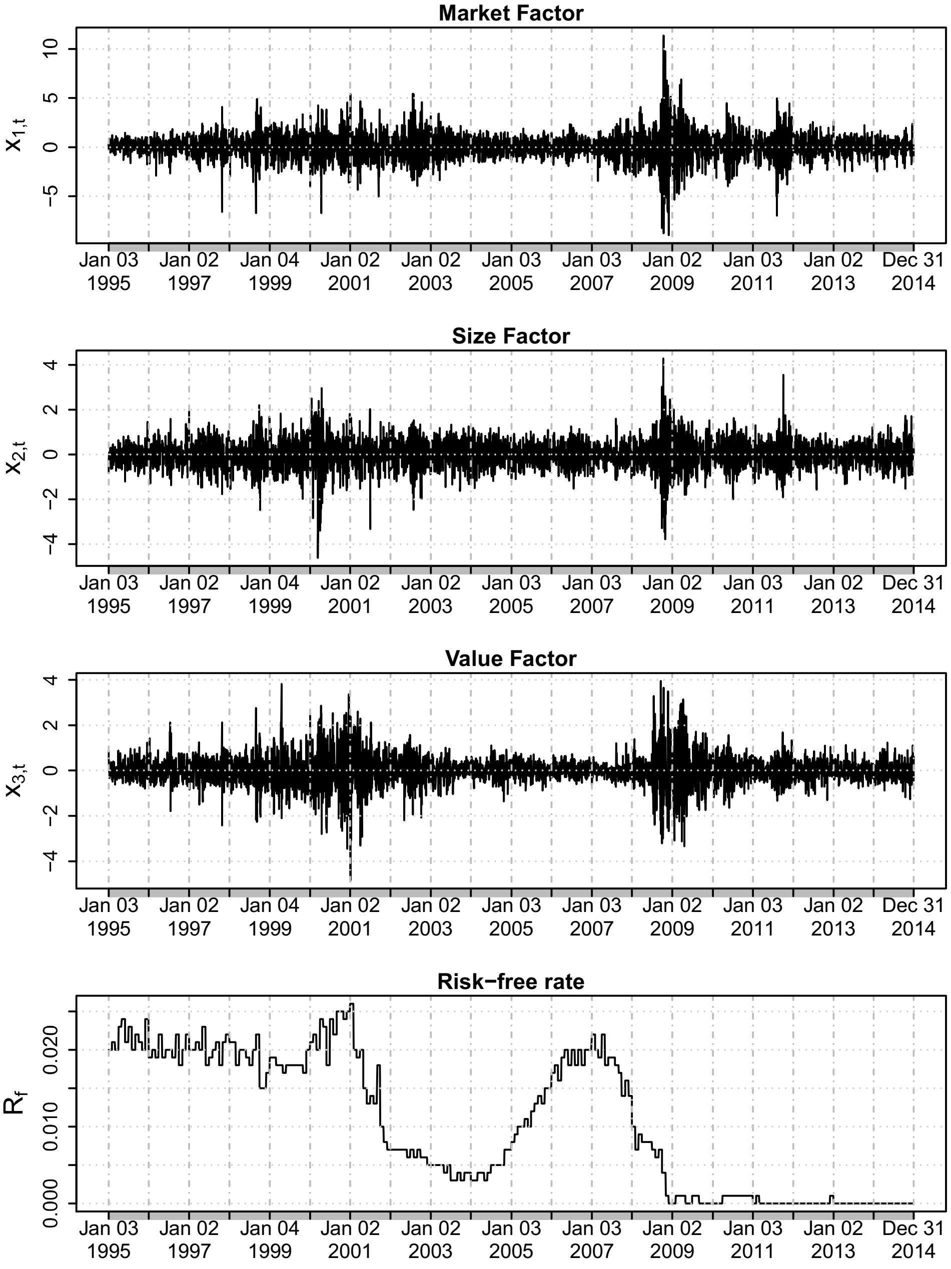
This table gives the name of 'Dataset j' followed by a brief description, quoted from Kenneth French's website. The fourth column is the value p_n once we remove components containing missing values.

Table 7.3: **Description of the 49 industry portfolios**

k	$y_{k,t}$	Industry name	k	$y_{k,t}$	Industry name
1	`Agric`	Agriculture	26	`Guns`	Defense
2	`Food`	Food Products	27	`Gold`	Precious Metals
3	`Soda`	Candy & Soda	28	`Mines`	Non-Metallic & Industrial Metal Mining
4	`Beer`	Beer & Liquor	29	`Coal`	Coal
5	`Smoke`	Tobacco Products	30	`Oil`	Petroleum and Natural Gas
6	`Toys`	Recreation	31	`Util`	Utilities
7	`Fun`	Entertainment	32	`Telcm`	Communication
8	`Books`	Printing & Publishing	33	`PerSv`	Personal Services
9	`Hshld`	Consumer Goods	34	`BusSv`	Business Services
10	`Clths`	Apparel	35	`Hardw`	Computers
11	`Hlth`	Healthcare	36	`Softw`	Computer Software
12	`MedEq`	Medical Equipment	37	`Chips`	Electronic Equipment
13	`Drugs`	Pharmaceutical Products	38	`LabEq`	Measuring and Control Equipment
14	`Chems`	Chemicals	39	`Paper`	Business Supplies
15	`Rubbr`	Rubber & Plastic Products	40	`Boxes`	Shipping Containers
16	`Txtls`	Textiles	41	`Trans`	Transportation
17	`BldMt`	Construction Materials	42	`Whlsl`	Wholesale
18	`Cnstr`	Construction	43	`Rtail`	Retail
19	`Steel`	Steel Works Etc	44	`Meals`	Restaurants, Hotels, Motels
20	`FabPr`	Fabricated Products	45	`Banks`	Banking
21	`Mach`	Machinery	46	`Insur`	Insurance
22	`ElcEq`	Electrical Equipment	47	`RlEst`	Real Estate
23	`Autos`	Automobiles and Trucks	48	`Fin`	Trading
24	`Aero`	Aircraft	49	`Other`	Almost Nothing
25	`Ships`	Shipbuilding, Railroad Equipment			

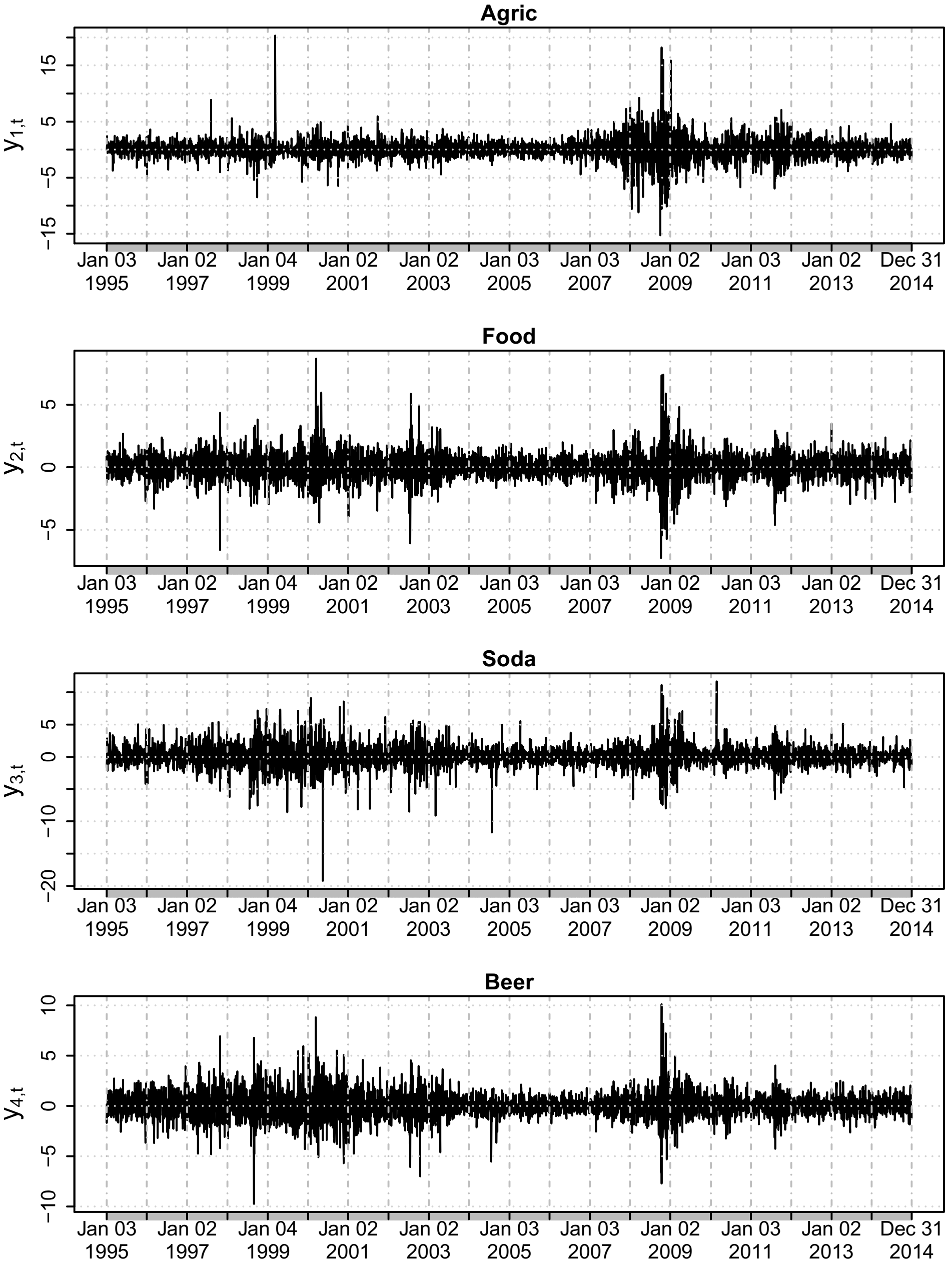
This table gives the labelling and a brief description of industrial sectors which form the 49 Industry Portfolios dataset. Precise details of their construction are given on Kenneth French's website.

Figure 30: Returns plot of factors and risk free rate



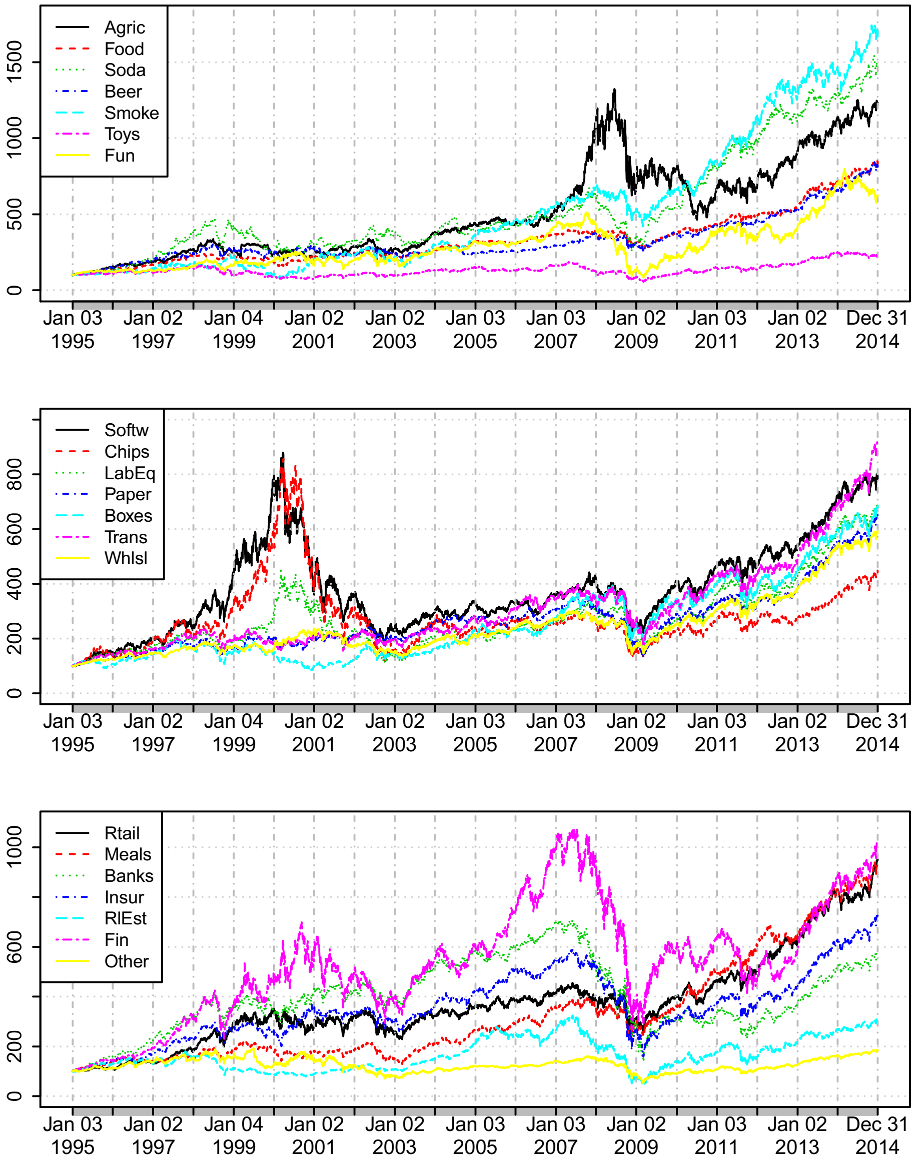
Returns plots of the (daily) three Fama-French factors, along with the (daily) risk-free rate of return.

Figure 31: Returns plots of $y_{1,t}$, $\cdots$,$y_{4,t}$



This figure shows the (daily) returns plots for the first four industrial portfolios in Dataset 1 (Agriculture, Food Products, Candy & Soda, and Beer & Liquor)

Figure 32: Comparison of industrial portfolios



This figure shows the (daily) cumulative return with a starting balance of £100 for some of the industrial portfolios from Dataset 1.

7.2 A trading strategy example

In this section, we compare the three portfolio allocations, ($\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$), along with the market portfolio, year by year from Jan 3rd 1995 to Dec 31st 2014 using a simple trading strategy. For each year we trade on each trading day, which is approximately $T = 252$ trading days per year. At the beginning of each year we assume we have an initial balance of £100. Although this initial choice is arbitrary, it is a useful way of comparing the performance during the course of a year. We assume no transaction costs, allow for short selling, and assume that all possible portfolio allocations are attainable. Our trading strategy consists of forming a portfolio allocation $\hat{\mathbf{w}}$ the end of each trading day and holding it until the end of the next trading day. Between day $t-1$ and day t, we obtain the portfolio return

$$R_t(\hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}} Y_t + R_{f,t}$$

where $\hat{\mathbf{w}}$ is formed based on $(X_{t-j}^{\mathrm{T}},\ Y_{t-j}^{\mathrm{T}})$, $j = 1, \cdots, n$, for some look-back integer n, and $R_{f,t}$ is the risk-free rate. In this section, we set $n = 500$ and $\delta = 1\%$.

We use Dataset 1 (49 Industry Portfolios) to demonstrate how the trading strategy works. We plot the balances at the end of each trading day over the course of each year in Figures 33 - 37. From these figures alone, it is clear that $\hat{\mathbf{w}}_{\text{FACE}}$ performs significantly better than the other three in terms of cumulative return.

We remark that although $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ are all constructed based on Markowitz's formula, the difference between them lies in the way to estimate the covariance matrix of returns, which appears in Markowitz's formula. Both $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ do not take into account the dynamic feature of the covariance matrix in their estimation, but $\hat{\mathbf{w}}_{\text{FACE}}$ does. This is the fundamental reason why $\hat{\mathbf{w}}_{\text{FACE}}$ performs significantly better than $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$.

One may argue that if $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ used fewer observations in their moving window to estimate the covariance matrix they would start to take the dynamic feature into account, potentially improving their performance. This issue is addressed in Section 7.4 when we explore alternative choices for n. We will see that even if $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ only use the observations in a carefully chosen moving window, $\hat{\mathbf{w}}_{\text{FACE}}$ still outperforms them.

To have a tangible idea about whether the covariance matrix is dynamic or not, we plot the estimated intercept and coefficients of $x_{1,t}$, $x_{2,t}$ and $x_{3,t}$, interpreted as the impact of the factors, for each of the first four components of Y_t in Figure 38. One can see that these coefficients are dynamic rather than constant, which implies the covariance matrix is also dynamic.

It is interesting to have a closer look at the performances of the four strategies in the volatile time period 2007-2009 during which the financial crisis took place (see Figure 36). During 2007, $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ all perform reasonably well, with $\hat{\mathbf{w}}_{\text{FACE}}$ slightly better. The market does not make much profit, and is beaten by the other three. In 2008, $\hat{\mathbf{w}}_{\text{FACE}}$ continuously does well whilst the other three do not make profit at all. In 2009, although $\hat{\mathbf{w}}_{\text{FACE}}$ does not do very well during some time periods, it adapts to the market change quickly and almost breaks even. The reason that $\hat{\mathbf{w}}_{\text{FACE}}$ can adapt to market change quickly is because it takes into account the dynamic feature of the covariance matrix of returns. On the other hand, both $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ do very poorly, and in fact they almost lose all their money at the end of the year. In 2009, the market performs best, but still with very little profit.

A similar incident occured during the beginning of 2001, around the time of the dot-com bubble bursting. We see that $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ suffered huge losses right at the start. Although $\hat{\mathbf{w}}_{\text{FACE}}$ also suffered some losses at the start, it recovered by the end of the year with an overall annual return of 40%.

Figure 33: Daily balance during 1995-1998

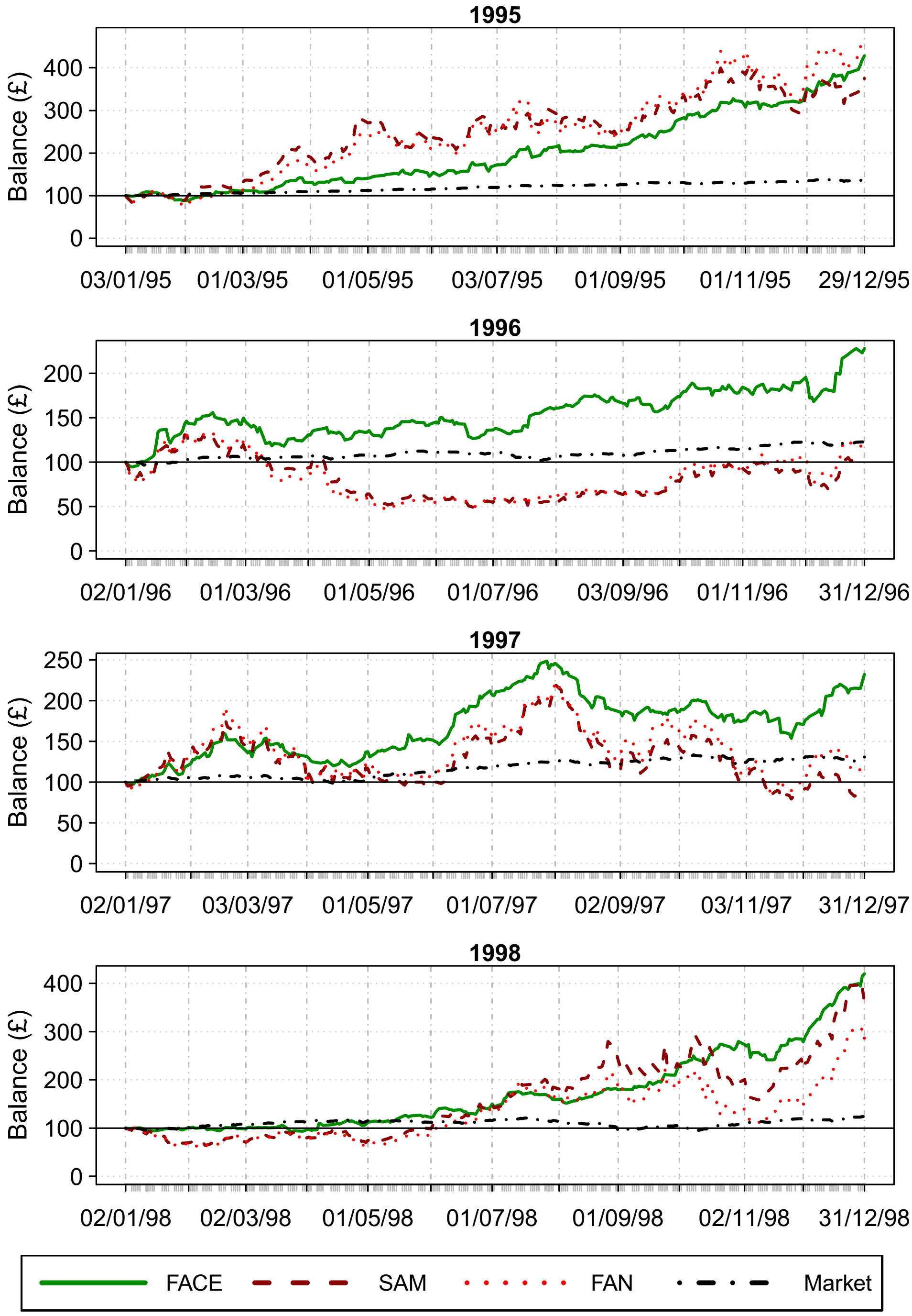


Figure 34: Daily balance during 1999-2002

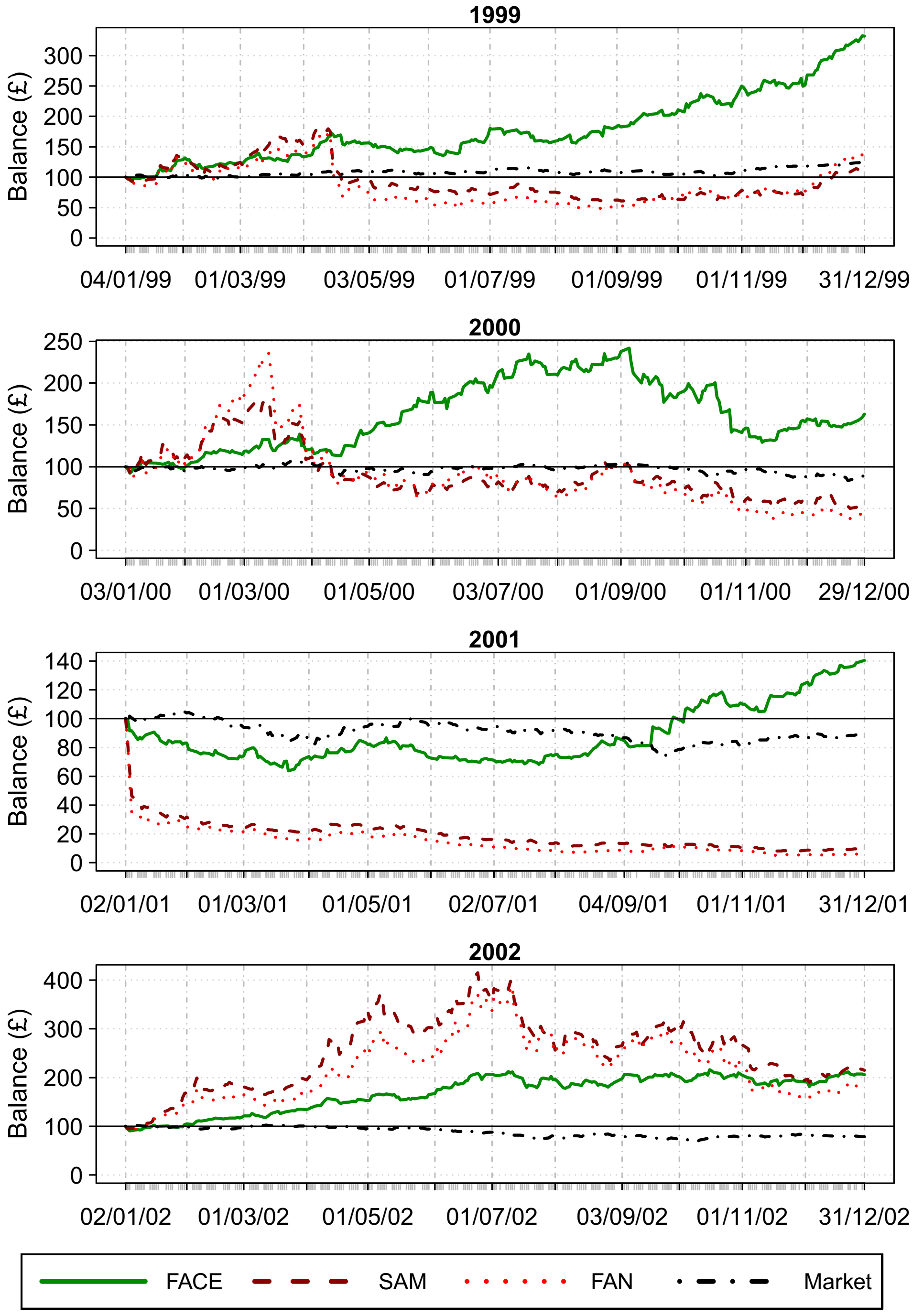


Figure 35: Daily balance during 2003-2006

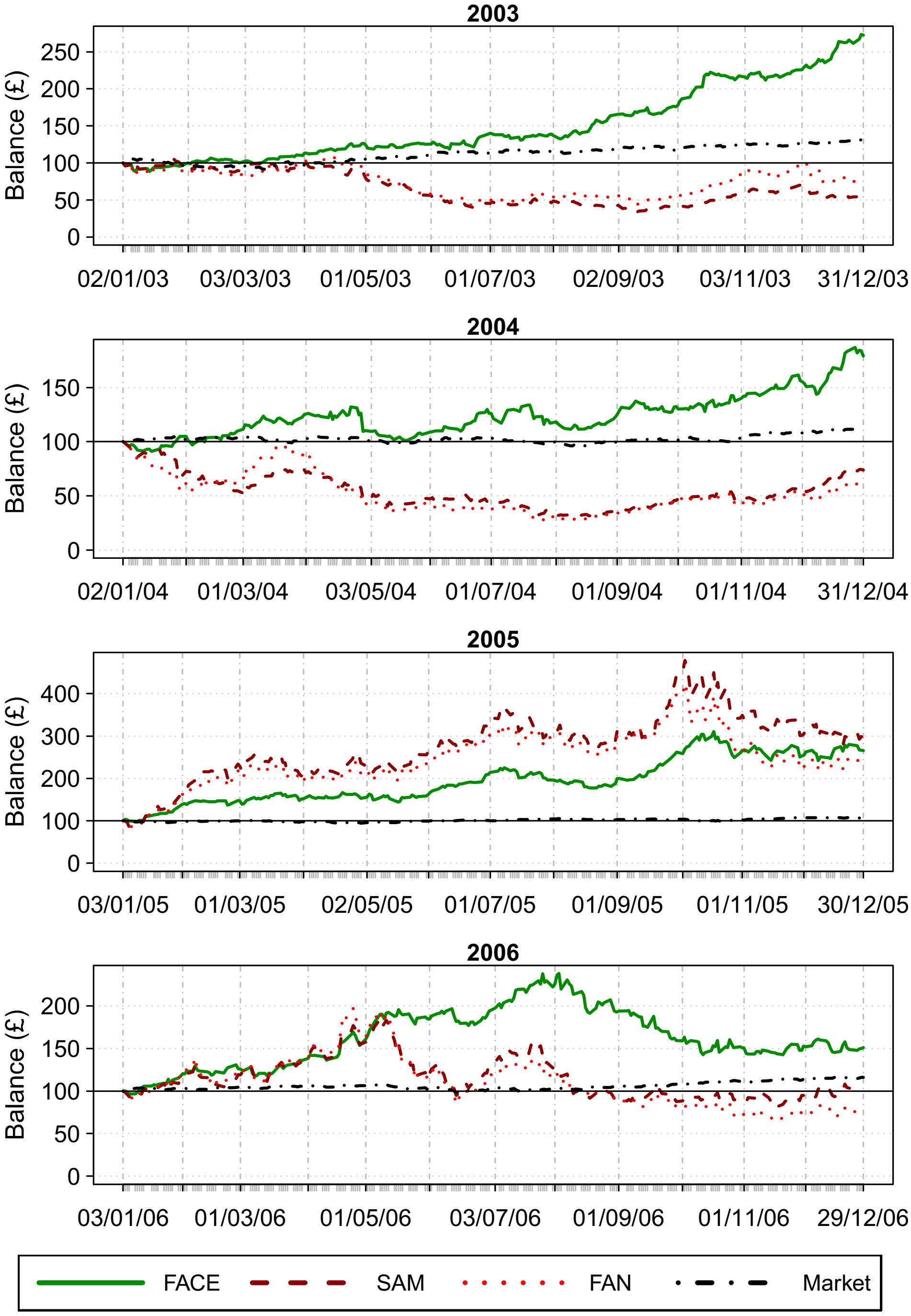


Figure 36: Daily balance during 2007-2010

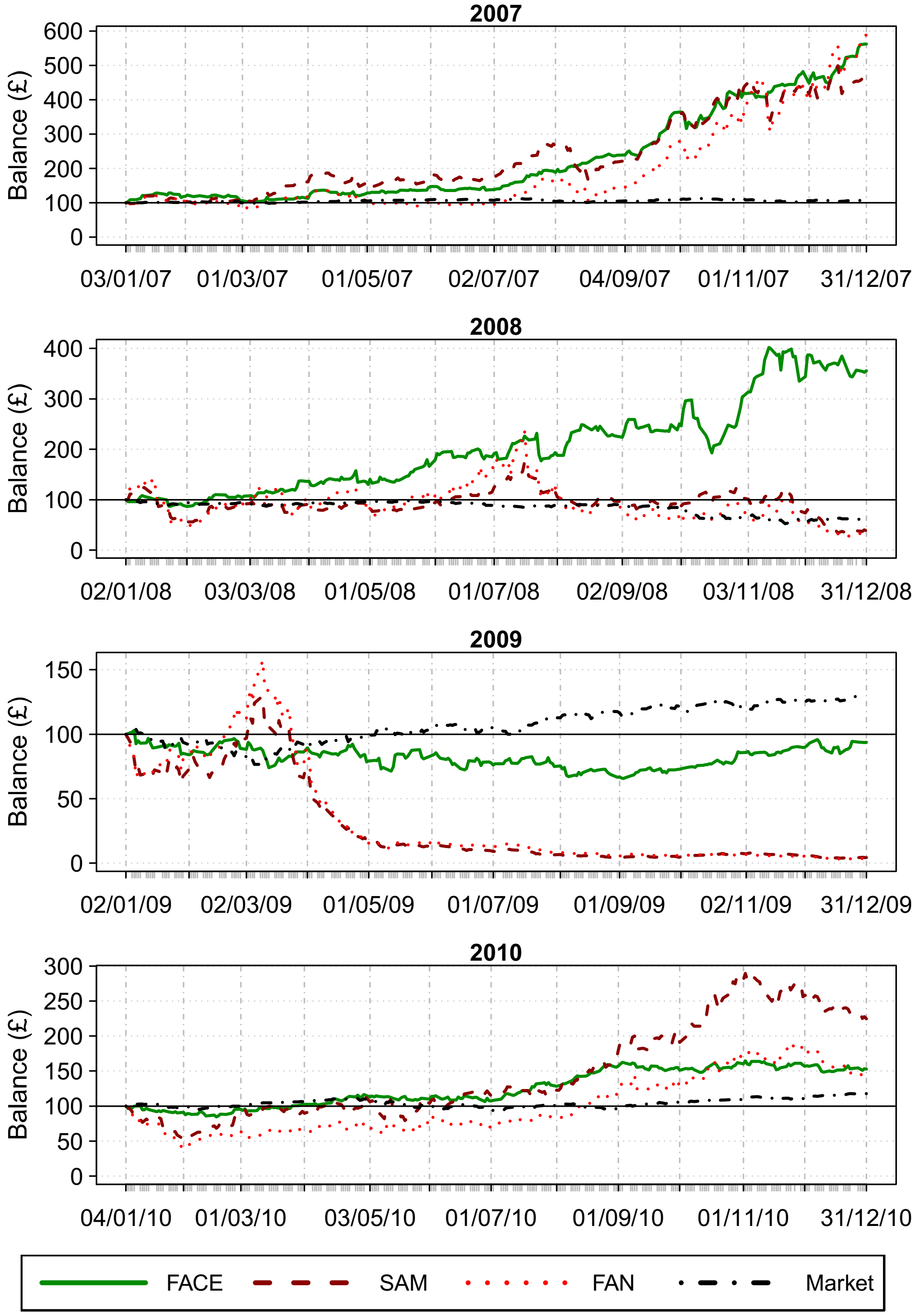


Figure 37: Daily balance during 2011-2014

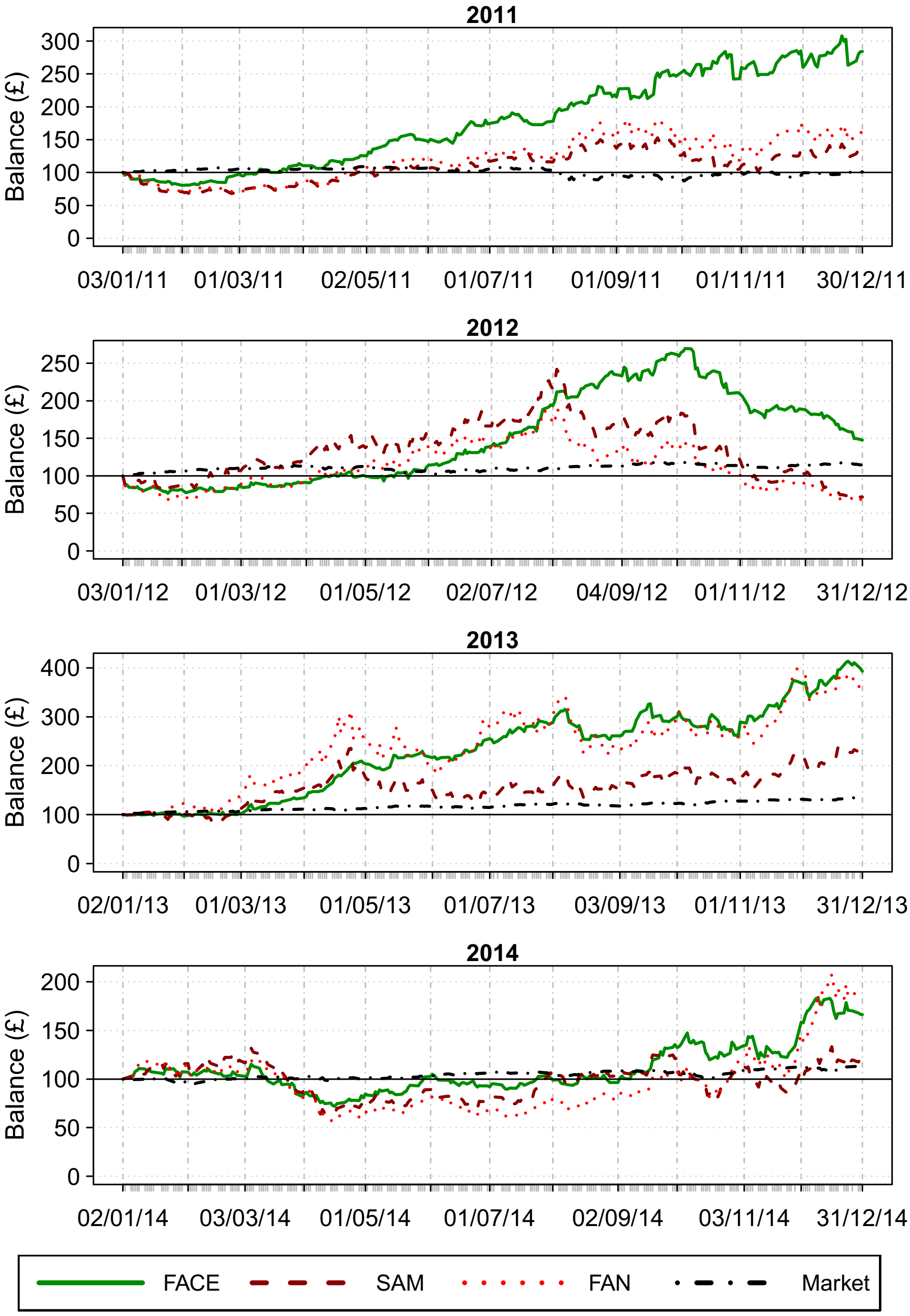
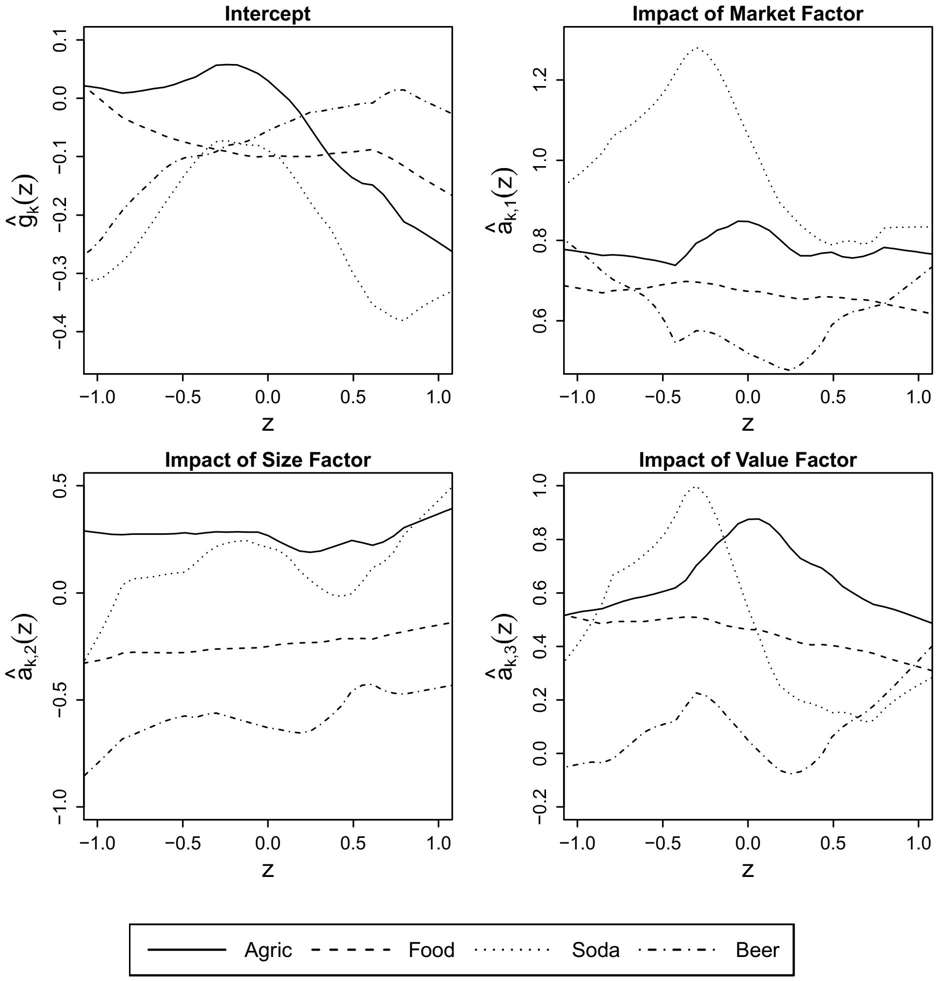


Figure 38: Coefficient functions of industry portfolios 1-4



This figure shows the estimated intercept and coefficient functions for the market, size and value factors, for the first four industry portfolios (Agriculture, Food Products, Candy & Soda, and Beer & Liquor) on the first day of trading.

7.3 Annualized Sharpe Ratio

In Section 7.2, we saw evidence to suggest that the proposed portfolio allocation $\hat{\mathbf{w}}_{\text{FACE}}$ can outperform $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ well when applied to Dataset 1. Assessing the performance of any trading strategy can be accurately, and more concisely, measured by using the Sharpe ratio, annualized to take into account the number of trading days in the year. With the realised returns $R_t(\hat{\mathbf{w}})$, $t = 1, \cdots, T$, we can calculate the annualized Sharpe ratio

$$\text{SR}(\hat{\mathbf{w}}) = \frac{\bar{R}(\hat{\mathbf{w}})}{\text{SD}(\hat{\mathbf{w}})}\sqrt{T},$$

where

$$\bar{R}(\hat{\mathbf{w}}) = \frac{1}{T}\sum_{t=1}^{T}\left\{R_t(\hat{\mathbf{w}}) - R_{f,t}\right\},$$

$$\text{SD}(\hat{\mathbf{w}}) = \left[\frac{1}{T}\sum_{t=1}^{T}\left\{R_t(\hat{\mathbf{w}}) - R_{f,t} - \bar{R}(\hat{\mathbf{w}})\right\}^2\right]^{1/2},$$

and $R_{f,t}$ is the risk-free rate on day t.

Another commonly used measure that practitioners use to measure the risk of a trading strategy is the maximum drawdown metric

$$\text{MDD}(\hat{\mathbf{w}}) = \max_{t\in[0,T]}\left[\max_{s\in[0,t]}\xi(s) - \xi(t)\right]$$

where $\xi(t)$ is the cumulative return process. This can be intuitively understood as the maximum historic decline over the period of interest, and is a useful way of measuring the risk associated with a portfolio allocation.

Using the same trading strategy as described in Section 7.2, and still using $n = 500$ with $\delta = 1\%$, we compute $\text{SR}(\hat{\mathbf{w}})$ and $\text{MDD}(\hat{\mathbf{w}})$

for each year from 1995 to 2014 and for each dataset given in Table 7.2. From the results given in Figure 39, we see that the Sharpe ratios for $\hat{\mathbf{w}}_{\text{FACE}}$ are consistently higher than that of $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$. Also, from Figure 40 we see that the maximum drawdowns for $\hat{\mathbf{w}}_{\text{FACE}}$ are consistently smaller than those of $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$. In a lot of cases, $\hat{\mathbf{w}}_{\text{FACE}}$'s maximum drawdowns are similar to that of the market, but with much higher Sharpe ratios. These conclusions can all be equally found for Datasets 1-4, and hence provides more convincing evidence to support the conclusions from the previous section and the need for the proposed dynamic structure to estimate covariance matrices.

The striking effects of the financial crisis in 2008-2009 and the dot-com bubble bursting in the early 2000's can be best illustrated by looking at the maximum drawdowns. We see in both time periods the allocations $\hat{\mathbf{w}}_{\text{FAN}}$ and $\hat{\mathbf{w}}_{\text{SAM}}$ suffered huge losses in these time periods, whereas $\hat{\mathbf{w}}_{\text{FACE}}$ did not. In 2009, across all four datasets, the Sharpe ratios for $\hat{\mathbf{w}}_{\text{FAN}}$ and $\hat{\mathbf{w}}_{\text{SAM}}$ range from 0 to -3 whereas with $\hat{\mathbf{w}}_{\text{FACE}}$ they range from 0 to 2. This evidence suggests how $\hat{\mathbf{w}}_{\text{FACE}}$ is more robust to areas of financial volatility thanks to the way it adapts to market change.

Figure 39: Sharpe ratios with $n = 500$ and $\delta = 1.0\%$.

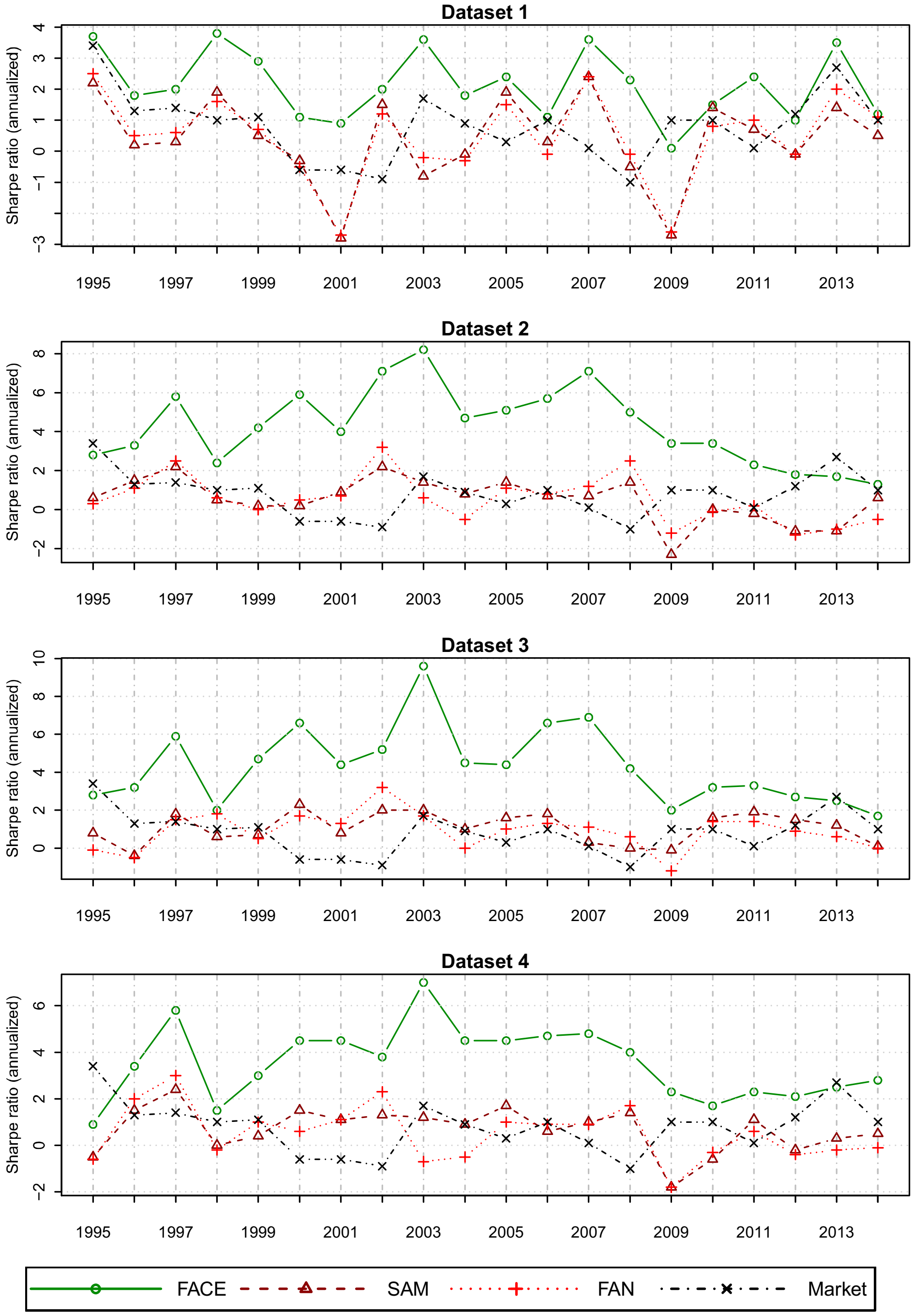
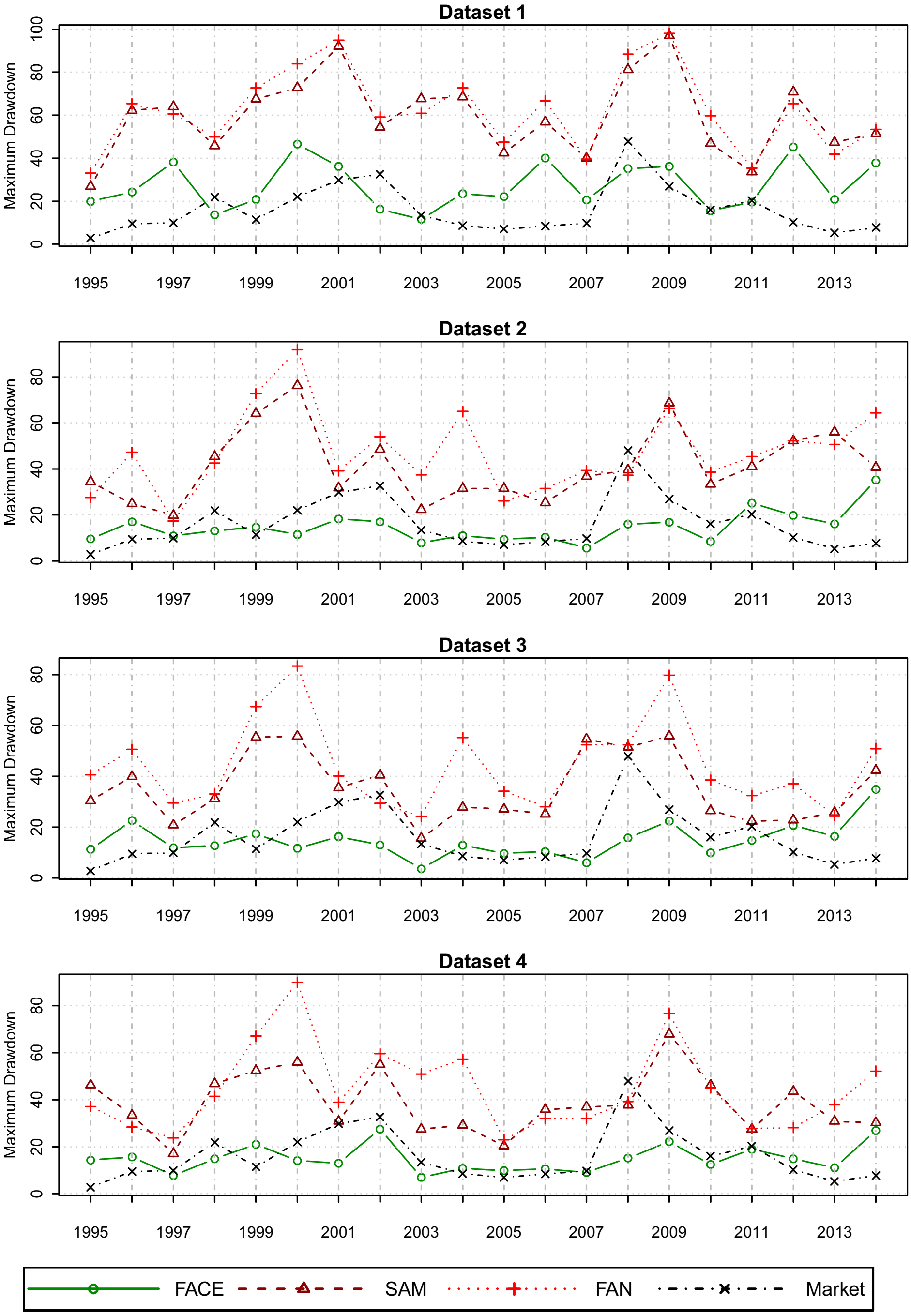


Figure 40: Maximum Drawdowns with $n = 500$ and $\delta = 1.0\%$.



7.4 Sensitivity to choice of sample size

A crucial question is whether the performance $\hat{\mathbf{w}}_{\text{FACE}}$, in terms of (cumulative) portfolio returns, is sensitive to the choice of sample size n used. In addition to this, one may argue that if $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ used fewer observations in their moving window to estimate the covariance matrix they would start to take the dynamic feature into account, potentially improving their performance.

In this section, we repeat the backtesting experiment described in Section 7.2 using $n = 300$ and $n = 100$ to estimate the covariance matrix. This way, the portfolios formed would only take into account the most recent observations, and would react better to sudden changes in the market. In the same way as before, we trade on each trading day and assume an initial balance of £100 at the beginning of each year. However, for brevity, we only report the balance at the end of the last trading day of each year. This can also be interpreted as a cumulative return throughout the year. We do this, independently, for Datasets 1-4 and report the results in Tables 7.4 - 7.7 respectively. By looking at the balance at the end of each year, we can see that $\hat{\mathbf{w}}_{\text{FACE}}$ usually significantly outperforms $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$, regardless of whether $n = 100$, 300 or 500 is used.

By focusing on the effects of the financial crash in 2009, we do indeed see a small improvement for $\hat{\mathbf{w}}_{\text{FAN}}$ when a smaller sample size is used. For example, in Dataset 1, the final balances of $\hat{\mathbf{w}}_{\text{FAN}}$ are £68, £5 and £3, when $n = 100$, 300 and 500 respectively. Similarly, for $\hat{\mathbf{w}}_{\text{SAM}}$, they are £50, £9 and £4. Indeed, although there is an improvement when $n = 100$, the losses are still devastating due to the poor estimation of the covariance matrix. On the other hand, $\hat{\mathbf{w}}_{\text{FACE}}$ earns £94, £189 and £149 as a consequence of always acknowledging the dynamic feature. Datasets 2-4 also tell a similar story in that

$\hat{\mathbf{w}}_{\text{FACE}}$ performs well regardless of n. In these datasets, we also see multiple occurrences of $\hat{\mathbf{w}}_{\text{SAM}}$ losing all its money if too small a sample size is chosen.

The annualized Sharpe ratios for each dataset can be found in Figure 41 and Figure 42 for $n = 100$ and $n = 300$ respectively, and very similar conclusions can be made.

This evidence suggests that even if one carefully selects a sample size, attempting to capture the dynamics of the market, both $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ still perform poorly relative to $\hat{\mathbf{w}}_{\text{FACE}}$. Further to this, we see that $\hat{\mathbf{w}}_{\text{FACE}}$ is fairly robust to the choice of n. This is important because it shows $\hat{\mathbf{w}}_{\text{FACE}}$ did not perform well due to chance alone.

Table 7.4: Balances for Dataset 1 with different n

Year	$n = 100$			$n = 300$			$n = 500$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	229	172	222	554	275	355	428	375	475
1996	163	102	98	193	57	75	228	102	121
1997	177	138	155	298	143	205	232	94	125
1998	175	79	135	320	339	303	420	361	285
1999	120	59	77	253	115	167	332	114	131
2000	176	101	132	253	157	122	163	55	43
2001	128	53	60	166	50	49	140	10	6
2002	162	74	69	224	149	143	206	215	180
2003	162	57	99	135	41	47	273	54	75
2004	111	67	95	131	54	56	179	74	62
2005	183	200	171	186	161	153	265	293	236
2006	149	119	122	184	114	96	151	103	77
2007	236	191	236	398	319	339	563	472	590
2008	142	73	105	201	80	117	356	38	33
2009	149	50	68	189	9	5	94	4	3
2010	129	109	102	107	172	148	153	224	143
2011	180	110	96	192	94	126	284	134	161
2012	159	118	96	125	62	85	148	72	69
2013	233	197	229	417	180	277	393	229	371
2014	161	138	137	158	119	137	166	118	186

For Dataset 1, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$*,* $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$*'s annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for sample sizes* $n =$ 100, 300, 500.

Table 7.5: Balances for Dataset 2 with different n

Year	$n = 100$			$n = 300$			$n = 500$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	144	81	106	150	114	91	216	124	113
1996	197	237	162	255	195	161	291	201	171
1997	185	121	164	433	292	321	457	233	274
1998	196	83	92	194	91	115	190	120	129
1999	184	6	107	356	131	136	471	89	71
2000	475	4	229	898	188	263	991	98	94
2001	173	26	81	327	85	92	367	134	133
2002	280	9	117	576	255	294	477	207	349
2003	339	109	212	518	82	118	608	153	116
2004	311	71	142	260	165	111	285	126	68
2005	179	7	107	345	206	174	398	164	149
2006	181	41	139	352	133	146	483	129	132
2007	321	466	213	531	165	143	657	134	181
2008	225	3	155	393	169	364	602	206	517
2009	220	95	115	318	76	49	279	36	53
2010	193	205	98	186	66	111	208	95	89
2011	153	130	95	232	69	114	201	84	98
2012	109	21	74	134	50	44	170	55	50
2013	166	54	104	196	118	92	160	55	58
2014	170	96	159	154	129	99	149	117	65

For Dataset 2, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$*,* $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$*'s annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for sample sizes* $n =$ 100, 300, 500.

Table 7.6: Balances for Dataset 3 with different n

Year	$n = 100$			$n = 300$			$n = 500$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	159	0	99	185	104	84	217	130	100
1996	191	34	167	284	142	141	261	83	78
1997	183	107	143	436	156	208	584	218	195
1998	159	8	112	164	103	152	185	135	238
1999	162	2	96	326	134	107	629	131	114
2000	386	17	229	746	326	436	1151	405	318
2001	183	9	87	334	143	127	384	131	162
2002	233	10	82	434	205	231	305	186	315
2003	348	144	229	571	155	178	676	202	160
2004	266	127	174	286	137	90	290	129	90
2005	178	16	104	373	215	200	352	186	146
2006	224	385	148	429	166	122	712	198	168
2007	307	97	181	569	109	142	728	107	161
2008	180	0	82	410	96	173	505	84	129
2009	198	34	95	207	65	35	194	87	41
2010	190	0	119	275	142	161	241	183	191
2011	206	11	145	280	219	201	267	206	182
2012	132	295	87	183	200	107	203	170	129
2013	149	18	108	192	158	126	193	151	118
2014	138	58	116	191	109	109	185	94	91

For Dataset 3, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$*,* $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$*'s annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for sample sizes* $n =$ 100, 300, 500.

Table 7.7: Balances for Dataset 4 with different n

Year	$n = 100$			$n = 300$			$n = 500$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	117	87	83	130	61	77	129	78	75
1996	194	1144	211	276	186	219	314	191	249
1997	188	15	155	436	165	275	494	245	323
1998	203	6	117	229	94	84	150	100	98
1999	156	0	129	247	116	181	325	104	163
2000	408	63	231	645	320	302	543	247	112
2001	180	3	130	279	122	150	365	141	146
2002	255	1	126	431	134	282	274	159	297
2003	343	213	228	436	159	126	504	143	70
2004	265	156	189	293	133	114	303	133	72
2005	202	38	100	364	115	155	427	176	142
2006	158	19	104	304	108	115	396	126	140
2007	269	16	169	322	152	128	443	156	144
2008	179	28	112	363	244	236	482	219	281
2009	149	0	71	183	31	30	203	38	39
2010	126	18	68	165	57	87	151	72	80
2011	174	385	165	355	128	231	206	157	128
2012	128	3	84	156	64	54	179	85	77
2013	165	138	129	218	136	146	205	103	86
2014	133	9	115	245	163	133	260	115	80

For Dataset 4, this table shows a comparison of Face, Sam and Fan's annual balance (rounded to the nearest pound), assuming an initial balance of £100 at the start of each year, for sample sizes $n = 100,\ 300,\ 500$.

Figure 41: Sharpe ratios with $n = 100$ and $\delta = 1\%$

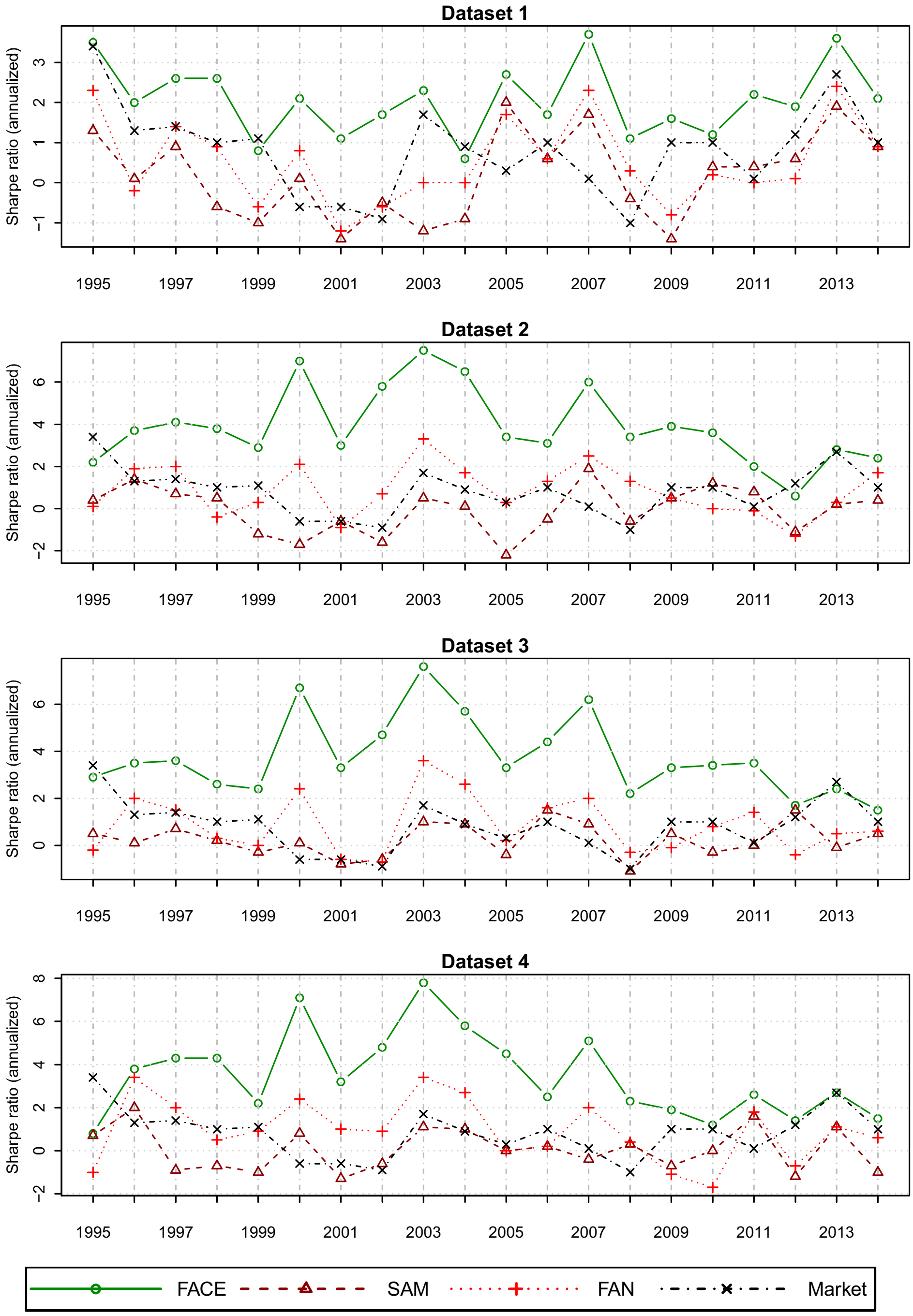
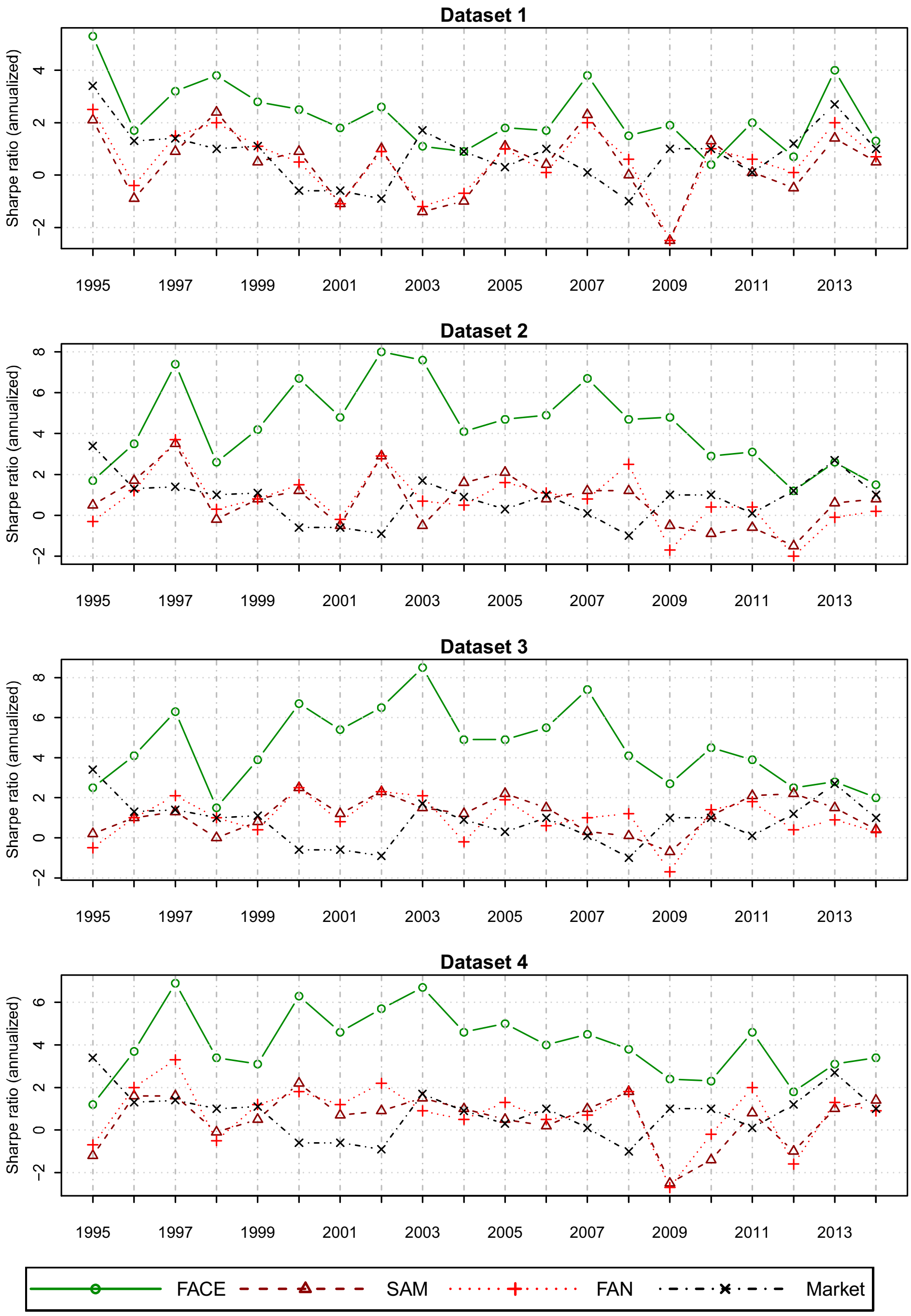


Figure 42: Sharpe ratios with $n = 300$ and $\delta = 1\%$



7.5 Sensitivity to choice of target return

In previous sections, we have found a significant amount of evidence for using the proposed dynamic structure. In particular, we saw that $\hat{\mathbf{w}}_{\text{FACE}}$ can outperform $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ over 20 years worth of data across four datasets, and that these conclusions hold for different choices of n. As previously mentioned, the key difference between $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ lies in the way to estimate the covariance matrix of returns, which appears in Markowitz's formula. Another necessary quantity which appears in Markowitz's formula is the target return rate δ, which we have previously set equal to 1.0% for simplicity. An important question is whether the choice of δ has any significant effect on the overall performance of the trading strategy, and this will be the topic explored in this section.

We remark that a larger choice of δ means that an investor would be willing to take more risk, in exchange for a larger expected return on the allocated portfolio. This choice can sometimes be thought of as a personal preference which is specific to the investor or fund, depending on their risk tolerance and the returns they desire. This, naturally, can change over time depending on the volatility of the financial markets.

Choosing δ is related to the amount of exposure in certain assets, facilitated by short-selling. In practice, there are limits on how much one can short-sell, and so many portfolio allocations obtained using either $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ could be unattainable in real life. We ignore this technicality in the interest of keeping things simple, because it is easy to add further constraints to Markowitz's formula in order to restrict short selling. In Section 7.6, we explore an example of this where we limit the gross exposure of a portfolio. The crucial point, is that the estimation of covariance matrices is vital to forming a portfolio allocation regardless of these additional constraints.

In this section, we explore the sensitivity of $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ to the choice of δ by repeating the backtesting experiment from Section 7.2 using $\delta = 0.5\%$ and $\delta = 1.5\%$ in addition to $\delta = 1.0\%$. We set $n = 500$ in order to keep the number of comparisons to a minimum. For these different target returns, Tables 7.8 - 7.11 show the balance on the final trading day of each year, assuming an initial balance of $\pounds 100$ at the start of each year. These results agree with the conclusions from the previous sections in the sense that $\hat{\mathbf{w}}_{\text{FACE}}$ performs better than $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ the vast majority of time in the 20 year period, and across all four datasets. We often see that as δ increases, the cumulative returns resulting from $\hat{\mathbf{w}}_{\text{FACE}}$ also increase, whereas with $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ they decrease. The reason why $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ decrease here is because they are often exposed heavily in certain assets and do not adapt to market change quickly at all. The increase in the portfolios volatility, associated with an increase in δ, also contributes to this decline. However, we note that $\hat{\mathbf{w}}_{\text{FACE}}$ can also lose money by an increase in δ, although only relatively small amounts compared to $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$. For example, during the financial crisis in 2009 for Dataset 1, $\hat{\mathbf{w}}_{\text{FACE}}$ obtains an annual balances of $\pounds 101$, $\pounds 94$, and $\pounds 83$ for $\delta = 0.5\%$, 1.0%, and 1.5% respectively. On the other hand, $\hat{\mathbf{w}}_{\text{SAM}}$ ended the year with $\pounds 24$, $\pounds 4$, and $\pounds 1$, and $\hat{\mathbf{w}}_{\text{FAN}}$ with $\pounds 20$, $\pounds 3$, and $\pounds 0$.

Figure 43 and Figure 44 show annualized Sharpe ratios for $\delta = 0.5\%$ and 1.5% respectively. Comparing these with Figure 39, we still see that $\hat{\mathbf{w}}_{\text{FACE}}$'s Sharpe ratios are consistently greater than zero and outperform $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$. Since the Sharpe ratio measures the expected reward to risk tradeoff, it is not too surprising that the plots for $\hat{\mathbf{w}}_{\text{FACE}}$ look fairly similar across each δ.

Table 7.8: Balances for Dataset 1 with different δ

Year	$\delta = 0.5\%$			$\delta = 1\%$			$\delta = 1.5\%$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	233	232	262	428	375	475	755	542	757
1996	163	115	124	228	102	121	304	79	103
1997	157	113	127	232	94	125	327	66	104
1998	198	198	179	420	361	285	856	558	377
1999	169	106	115	332	114	131	622	105	119
2000	137	81	75	163	55	43	182	31	19
2001	130	35	28	140	10	6	147	2	1
2002	144	151	137	206	215	180	287	274	209
2003	183	87	100	273	54	75	398	31	51
2004	145	103	95	179	74	62	214	47	35
2005	163	172	158	265	293	236	410	443	310
2006	139	117	100	151	103	77	157	82	52
2007	248	235	270	563	472	590	1212	833	1060
2008	180	63	71	356	38	33	635	17	8
2009	101	24	20	94	4	3	83	1	0
2010	121	170	129	153	224	143	189	260	136
2011	193	133	153	284	134	161	396	123	148
2012	128	92	95	148	72	69	163	49	43
2013	227	175	226	393	229	371	650	253	511
2014	146	129	161	166	118	186	176	95	181

For Dataset 1, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$*,* $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$*'s annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for target returns* $\delta = 0.5\%,\ 1\%,\ 1.5\%$*. The sample size used was* $n = 500$*.*

Table 7.9: Balances for Dataset 2 with different δ

Year	$\delta = 0.5\%$			$\delta = 1\%$			$\delta = 1.5\%$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	169	126	122	216	124	113	271	118	101
1996	198	168	157	291	201	171	419	227	174
1997	253	189	204	457	233	274	807	275	352
1998	137	102	107	190	120	129	259	135	148
1999	222	102	100	471	89	71	957	69	43
2000	380	123	128	991	98	94	2481	69	50
2001	210	138	133	367	134	133	626	126	125
2002	216	161	192	477	207	349	1036	257	600
2003	307	156	149	608	153	116	1183	147	86
2004	170	122	93	285	126	68	470	124	47
2005	207	139	128	398	164	149	749	187	166
2006	246	123	125	483	129	132	934	131	135
2007	247	121	130	657	134	181	1717	142	236
2008	195	121	184	602	206	517	1797	317	1250
2009	195	66	82	279	36	53	391	19	32
2010	146	105	101	208	95	89	292	82	75
2011	146	100	99	201	84	98	269	66	87
2012	143	86	80	170	55	50	197	33	29
2013	150	92	94	160	55	58	165	32	34
2014	128	114	88	149	117	65	167	112	45

For Dataset 2, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$'s annual balance (rounded to the nearest pound), assuming an initial balance of £100 at the start of each year, for target returns $\delta = 0.5\%$, 1%, 1.5%. The sample size used was $n = 500$.

Table 7.10: Balances for Dataset 3 with different δ

Year	$\delta = 0.5\%$			$\delta = 1\%$			$\delta = 1.5\%$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	168	132	115	217	130	100	273	124	83
1996	188	108	106	261	83	78	356	61	55
1997	285	181	174	584	218	195	1164	250	208
1998	130	108	140	185	135	238	255	159	377
1999	255	120	121	629	131	114	1478	131	95
2000	389	253	218	1151	405	318	3285	579	365
2001	225	135	153	384	131	162	639	123	164
2002	179	147	186	305	186	315	514	228	512
2003	313	174	165	676	202	160	1436	225	151
2004	180	128	109	290	129	90	457	126	71
2005	187	141	126	352	186	146	649	237	161
2006	299	157	145	712	198	168	1653	238	186
2007	244	103	121	728	107	161	2117	104	200
2008	189	79	89	505	84	129	1304	80	163
2009	156	104	69	194	87	41	236	68	22
2010	163	139	153	241	183	191	349	233	224
2011	166	150	140	267	206	182	418	270	221
2012	151	139	119	203	170	129	266	198	133
2013	160	141	127	193	151	118	229	157	107
2014	145	106	104	185	94	91	228	80	75

For Dataset 3, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$*,* $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$*'s annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for target returns* $\delta = 0.5\%,\ 1\%,\ 1.5\%$*. The sample size used was* $n = 500$*.*

Table 7.11: Balances for Dataset 4 with different δ

Year	$\delta = 0.5\%$			$\delta = 1\%$			$\delta = 1.5\%$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	133	102	102	129	78	75	122	57	53
1996	208	165	187	314	191	249	462	211	312
1997	264	191	218	494	245	323	905	301	458
1998	117	94	92	150	100	98	189	103	99
1999	175	107	143	325	104	163	576	90	154
2000	266	190	132	543	247	112	1072	281	69
2001	230	142	148	365	141	146	566	136	139
2002	180	145	193	274	159	297	410	166	421
2003	266	150	114	504	143	70	937	131	41
2004	184	134	99	303	133	72	488	126	49
2005	205	138	126	427	176	142	863	215	154
2006	221	126	132	396	126	140	693	122	142
2007	209	130	121	443	156	144	911	174	161
2008	174	117	134	482	219	281	1274	365	510
2009	155	69	70	203	38	39	260	20	20
2010	126	92	96	151	72	80	177	54	65
2011	152	135	120	206	157	128	271	168	125
2012	145	105	98	179	85	77	216	66	57
2013	164	118	110	205	103	86	251	86	64
2014	171	115	97	260	115	80	380	107	61

For Dataset 4, this table shows a comparison of $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$'s annual balance (rounded to the nearest pound), assuming an initial balance of £100 at the start of each year, for target returns $\delta = 0.5\%$, 1%, 1.5%. The sample size used was $n = 500$.

Figure 43: Sharpe ratios ($\delta = 0.5\%,\ n = 500$)

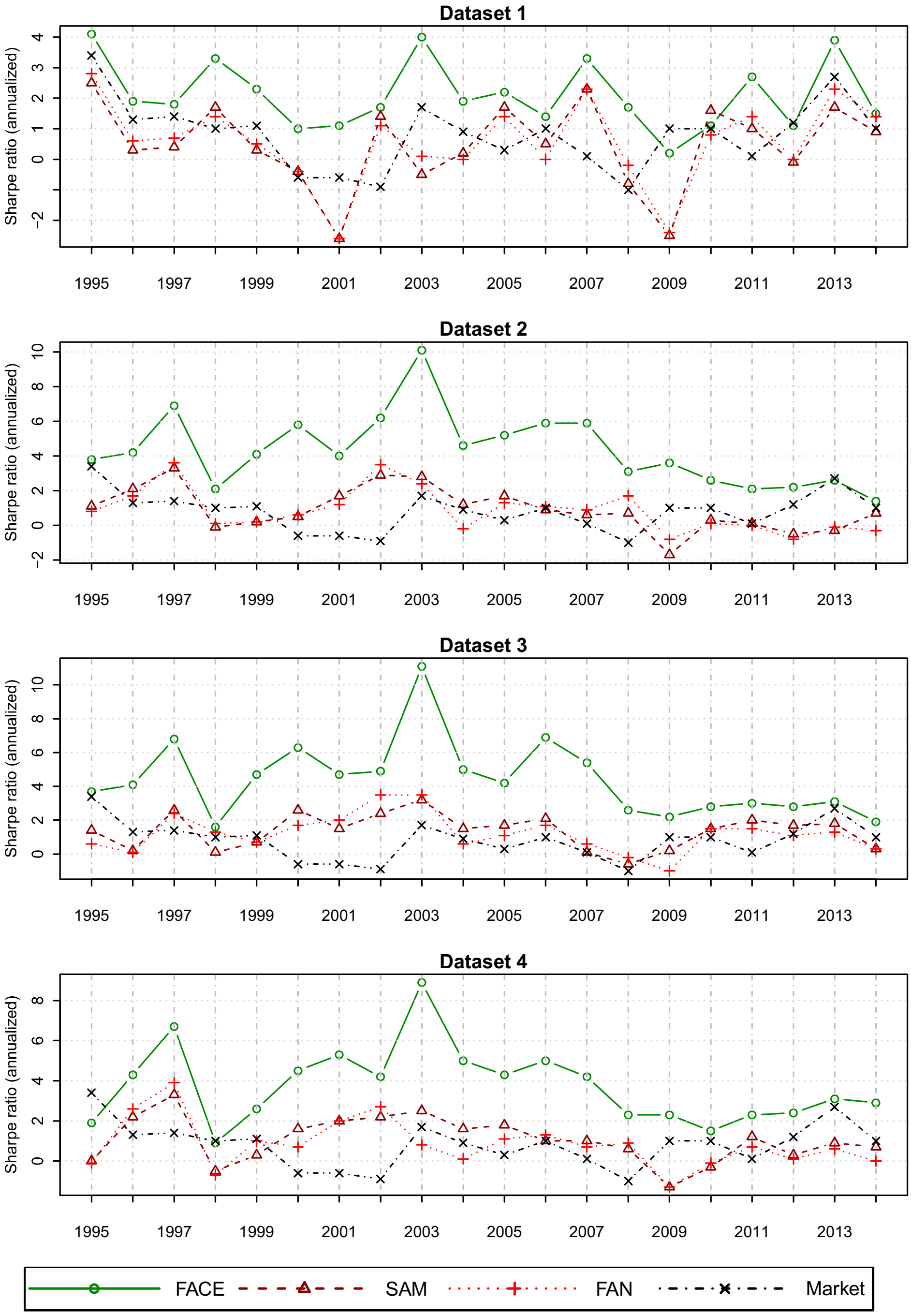
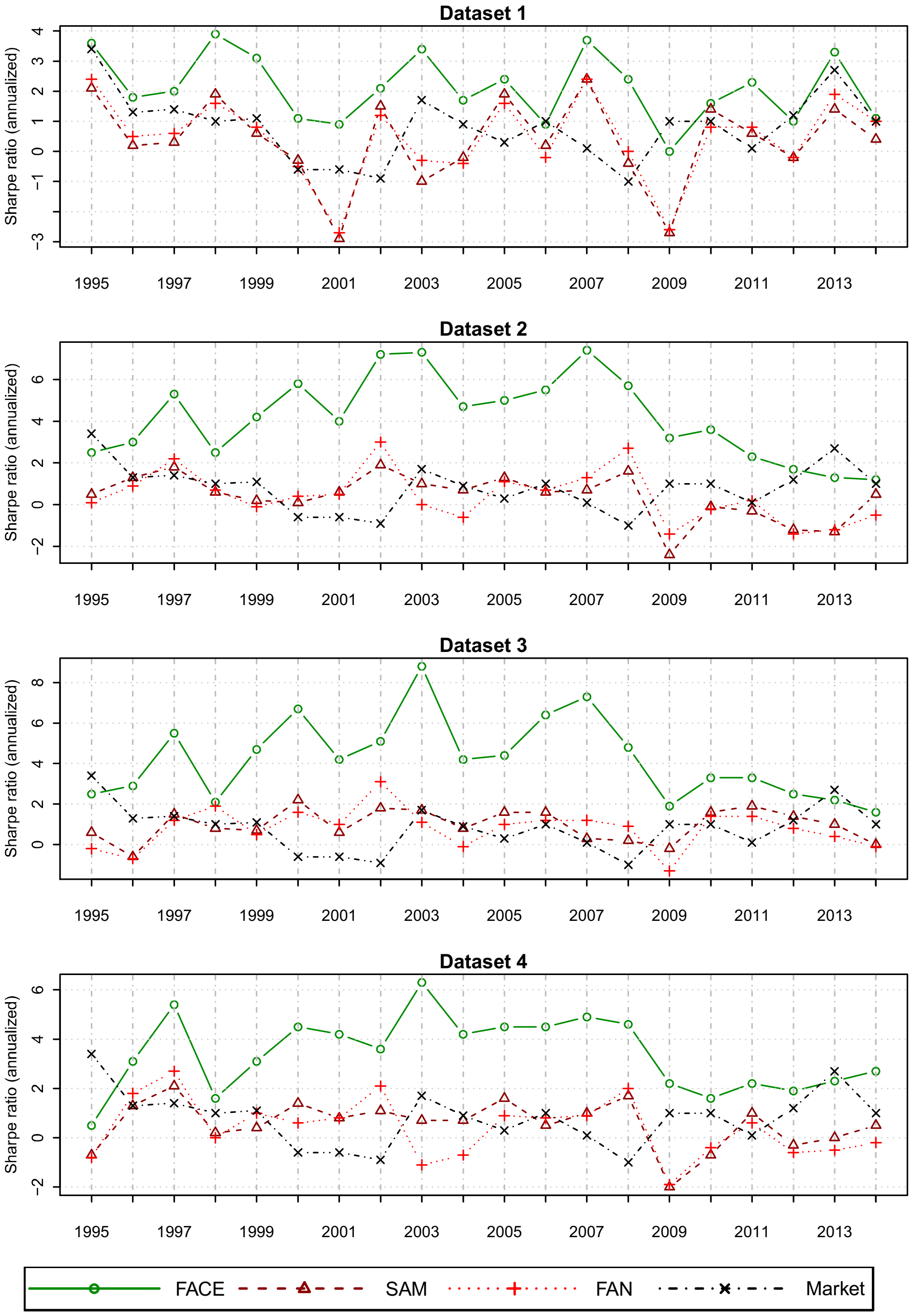


Figure 44: Sharpe ratios ($\delta = 1.5\%,\ n = 500$)



7.6 Gross-exposure constraints

In the thesis so far, we have explored portfolio allocations based on Markowitz's formula, which is the solution to the following constrained optimisation problem

$$\begin{aligned}&\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}\\&\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1}_{p_n} = 1 \quad \text{and} \quad \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \delta\end{aligned} \tag{7.1}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of excess asset returns and $\boldsymbol{\mu}$ is the expected vector of asset returns. As previously mentioned, the difference between $\hat{\mathbf{w}}_{\mathrm{FACE}}$, $\hat{\mathbf{w}}_{\mathrm{SAM}}$ and $\hat{\mathbf{w}}_{\mathrm{FAN}}$ is how we estimate $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. It is possible to extend the above optimisation problem to include additional constraints, and we shall explore one such example in this section.

Practitioners are often subject to regulations with regards to short selling and can be under various exposure constraints in order to limit their leverage. This motivates the following extension to Markowitz's constrained optimisation problem

$$\begin{aligned}&\min_{\mathbf{w}} \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}\\&\text{subject to } \mathbf{w}^{\mathrm{T}}\mathbf{1}_{p_n} = 1,\ \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \delta \text{ and } \|\mathbf{w}\|_1 \le c\end{aligned} \tag{7.2}$$

where c is some constant which controls the maximum gross exposure which the optimum portfolio is allowed to have. To intuitively see how the additional constraint $\|\mathbf{w}\|_1 \le c$ prevents extreme positions in the portfolio, we remark that $c = 1$ means that no short sales are allowed and $c = \infty$ means there is no constraint on short sales. In previous chapters we have studied the special case $c = \infty$ which is potentially unrealistic in real life. The total proportions of long and short positions are

$$w^{+} = \frac{\|\mathbf{w}\|_1 + 1}{2} \quad \text{and} \quad w^{-} = \frac{\|\mathbf{w}\|_1 - 1}{2},$$

respectively, since $w^{+} + w^{-} = \|\mathbf{w}\|_1$ and $w^{+} - w^{-} = 1$.

Unlike (7.1), there is not an analytic solution to (7.2) due to the additional constraint on gross exposure. This optimisation problem has been studied previously (see Fan et al., 2008b and Fan et al., 2012) where it was noted the optimum portfolio allocation can be solved numerically using quadratic programming. For implementation purposes, we use the `quadprog` package in R (see Turlach and Weingessel, 2013) which uses the dual method of Goldfarb and Idnani (1982) and Goldfarb and Idnani (1983). The `solve.QP` function numerically solves a constrained minimisation problem of the form

$$\begin{aligned} \min_{\mathbf{b}} \quad & \frac{1}{2}\mathbf{b}^{\mathrm{T}}\mathbf{D}\mathbf{b} - \mathbf{d}^{\mathrm{T}}\mathbf{b} && (7.3) \\ \text{such that} \quad & \mathbf{A}^{\mathrm{T}}\mathbf{b} \geq \mathbf{b}_0 && (7.4) \end{aligned}$$

where we define the first m_{eq} rows of $\mathbf{A}^{\mathrm{T}}\mathbf{b} \geq \mathbf{b}_0$ to be equalities instead of inequalities. We remark that m_{eq} is an integer which is to be specified in the `solve.QP` function. One approach for solving this is to declare a new vector $\mathbf{u} = (u_1, \cdots, u_{p_n})^{\mathrm{T}}$ such that $u_i = |w_i|$ so that we can impose $\mathbf{1}^{\mathrm{T}}\mathbf{u} \leq c$. To achieve this, we set $m_{eq} = 2$ and set

$$\mathbf{d} = \mathbf{0}_{p_n \times 1}, \quad \mathbf{b} = (\mathbf{w}^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}, \quad \mathbf{b}_0 = (1, \delta, \mathbf{0}_{1\times p_n}, \mathbf{0}_{1\times p_n}\mathbf{0}_{1\times p_n}, -c)^{\mathrm{T}},$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{1}_{p_n\times 1} & \boldsymbol{\mu} & I_{p_n} & -I_{p_n} & 0 \times I_{p_n} & \mathbf{0}_{p_n\times 1} \\ \mathbf{0}_{p_n\times 1} & \mathbf{0}_{p_n\times 1} & I_{p_n} & I_{p_n} & I_{p_n} & -1 \times \mathbf{1}_{p_n\times 1} \end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix} 2\boldsymbol{\Sigma} & \mathbf{0}_{p_n\times p_n} \\ \mathbf{0}_{p_n\times p_n} & \kappa I_{p_n} \end{pmatrix}$$

for some small κ, for example $\kappa = 0.1$. Then, minimising the objective function in (7.2) becomes equivalent to

$$\min_{\mathbf{w},\mathbf{u}} \left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{w} + \frac{\kappa}{2} \mathbf{u}^{\mathrm{T}} \mathbf{u} \right)$$

and the constraints in (7.4) become equivalent to

$$\mathbf{1}^{\mathrm{T}} \mathbf{w} = 1, \quad \boldsymbol{\mu}^{\mathrm{T}} \mathbf{w} = \delta, \quad \mathbf{w} + \mathbf{u} \geq \mathbf{0}_{p_n \times 1},$$

$$-\mathbf{w} + \mathbf{u} \geq \mathbf{0}_{p_n \times 1}, \quad \mathbf{u} \geq \mathbf{0}_{p_n \times 1}, \quad -\mathbf{1}^{\mathrm{T}} \mathbf{u} \geq -c.$$

We implement the above methodology in a similar way to the previous sections, over the period from Jan 3rd 1995 to Dec 31st 2014. At the beginning of each year we still assume we have an initial balance of £100. However, one should note that occasionally there are trading days for which there is no solution to (7.2). This would occur, for example, if there are no possible portfolios which can simultaneously expect to reach the target return whilst meeting its leverage constraints. Hence we make a minor modification to our trading strategy to deal with this problem. If a solution $\hat{\mathbf{w}}$ exists, we form the portfolio allocation $\hat{\mathbf{w}}$ the end of the trading day and hold it until the end of the next trading day exactly as before. If, on the other hand, no solution exists, we simply do not trade and have the zero vector as the portfolio allocation. In this chapter we set $\delta = 0.5\%$ in order to increase the chances of finding a solution to (7.2). We apply this trading strategy using $c = 5$, 10, and 15 to Datasets 1-4, and report the balances at the end of the final trading day of each year in Tables 7.12-7.15 respectively.

Compared with the analysis in Section 7.2 to Section 7.5, we see that $\hat{\mathbf{w}}_{\mathrm{SAM}}$ and $\hat{\mathbf{w}}_{\mathrm{FAN}}$ have made an improvement from the point of view that they did not lose all their money in any year, like they did during 2001 and 2009 in Table 7.4 and Table 7.8. The reason for this is because the portfolios $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ were no longer as heavily exposed in certain technology related industries (Computer Software, Electrical Equipment, and Measuring and Control Equipment) and financial related industries (Banking, Insurance and Trading) which suffered huge losses due to the dot-com bubble and the financial crisis. In some cases, especially when $c = 5$, we note that they did not trade for some periods of time, which naturally stopped them losing money in the first place.

Overall, one can see higher end of year balances for $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ when the allocations were formed by (7.2) instead of (7.1). This same phenomenon has also been identified by Jagannathan and Ma (2002), who explained in more detail why imposing restrictions, such as no-short sales, can actually outperform Markowitz's portfolio allocation. This work was built upon by Fan et al. (2008b) and Fan et al. (2012), who explained that no-short-sales portfolios can be improved upon even more by allowing some short positions using (7.2).

Despite these improvements, however, we still do see that $\hat{\mathbf{w}}_{\text{FACE}}$ had higher returns than both $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ for the vast majority of years and across all four datasets. This can perhaps best be seen from the Sharpe ratio plots given in Figures 45 - 47. Although the gap in performance is smaller, we see that the Sharpe ratios of $\hat{\mathbf{w}}_{\text{FACE}}$ are still consistently much greater than zero, whereas $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ still suffer from negative Sharpe ratios in a number of time periods due to the fact that they are failing to adapt to market change.

In order to examine how the leverage associated with these portfolio allocations varies over time, we plot the end-of-month and end-of-year total gross exposure $\|\mathbf{w}\|_1$, for Dataset 1, in Figure 48 and Figure 49 respectively. It is clear that when c is relatively small, such as $c = 5$, the allocations consistently use the maximum amount of leverage possible in order to reach their target returns. In particular, we see a notable number of trading days, with $c = 5$, where $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ did not trade at all because no solution to (7.2) could be found. We see that as c increases to 10 and 15, the total gross exposure for $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ does not change quickly over time, and they often spend a long periods of time in heavily exposed positions. On the other hand, $\hat{\mathbf{w}}_{\text{FACE}}$ constantly changes its exposure, as it adapts to market change by using the dynamic structure to estimate the covariance matrix. It is precisely this reason which causes $\hat{\mathbf{w}}_{\text{FACE}}$ to outperform $\hat{\mathbf{w}}_{\text{SAM}}$ and $\hat{\mathbf{w}}_{\text{FAN}}$ in this context.

Table 7.12: Balances for Dataset 1 with $c = 5,\ 10,\ 15$.

Year	$c = 5$			$c = 10$			$c = 15$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	179	92	92	231	295	331	223	245	266
1996	136	89	88	161	114	118	163	115	122
1997	174	192	199	151	141	150	155	106	122
1998	199	131	113	194	188	172	193	193	175
1999	183	160	177	169	113	117	169	108	117
2000	133	58	63	138	80	74	138	80	74
2001	134	53	47	130	35	28	130	35	28
2002	146	124	117	144	151	136	144	151	136
2003	164	146	158	182	98	102	183	87	99
2004	149	108	105	144	111	113	144	110	97
2005	165	214	234	161	174	164	163	169	157
2006	92	64	65	134	109	109	141	111	103
2007	244	277	300	260	256	285	248	238	270
2008	158	55	58	186	65	69	182	64	69
2009	107	46	42	101	29	19	101	25	20
2010	117	143	129	119	175	137	119	170	137
2011	175	109	105	192	138	155	194	134	152
2012	134	124	115	127	79	93	129	85	95
2013	256	397	391	232	259	265	229	190	233
2014	139	157	190	149	145	170	146	136	165

For Dataset 1, this table shows a comparison of the three trading strategies annual balance (rounded to the nearest pound), assuming an initial balance of £100 at the start of each year, for $\delta = 0.5\%$ and for gross exposure constraints $c = 5,\ 10,$ and 15. The sample size used was $n = 500$.

Table 7.13: Balances for Dataset 2 with $c = 5,\ 10,\ 15$.

Year	$c = 5$			$c = 10$			$c = 15$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	170	100	100	174	135	133	171	126	121
1996	213	100	100	206	147	146	202	174	161
1997	259	120	121	264	252	244	255	204	211
1998	108	161	155	131	118	118	136	107	108
1999	252	131	138	220	76	89	225	84	100
2000	477	189	175	386	164	136	387	140	131
2001	199	129	139	208	135	133	208	137	133
2002	213	184	188	215	166	191	215	162	191
2003	298	182	186	301	171	153	302	157	148
2004	167	181	178	172	116	102	171	117	92
2005	203	142	135	209	142	129	208	144	133
2006	247	100	100	249	126	128	236	128	121
2007	185	100	100	252	70	76	249	102	130
2008	142	61	59	186	151	176	193	137	190
2009	206	90	91	198	80	80	197	68	84
2010	151	132	129	149	112	105	150	108	102
2011	122	95	97	133	98	114	142	109	106
2012	124	100	100	126	66	60	141	79	76
2013	158	102	102	156	128	121	152	100	95
2014	113	100	100	133	83	77	131	106	88

For Dataset 2, this table shows a comparison of the three trading strategies annual balance (rounded to the nearest pound), assuming an initial balance of £100 at the start of each year, for $\delta = 0.5\%$ and for gross exposure constraints $c = 5,\ 10,$ and 15. The sample size used was $n = 500$.

Table 7.14: Balances for Dataset 3 with $c = 5,\ 10,\ 15$.

Year	$c = 5$			$c = 10$			$c = 15$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	120	100	100	164	137	130	168	130	115
1996	157	100	100	183	119	113	189	124	116
1997	255	100	100	281	181	172	281	172	173
1998	117	269	265	135	150	148	129	120	141
1999	300	154	157	263	122	132	258	128	124
2000	428	440	401	383	284	237	391	249	219
2001	223	160	159	225	138	152	225	136	152
2002	181	170	174	178	151	185	178	150	185
2003	283	219	223	311	175	176	311	168	168
2004	173	127	127	178	122	108	179	129	109
2005	185	87	87	204	131	122	190	128	126
2006	214	100	100	308	144	145	294	152	141
2007	164	100	100	263	134	135	250	115	116
2008	142	100	100	198	108	114	192	91	93
2009	142	100	100	156	85	69	155	92	68
2010	147	106	108	164	145	146	162	145	150
2011	140	87	87	152	146	139	161	148	143
2012	139	100	100	162	161	158	154	135	126
2013	161	101	101	161	115	108	163	139	129
2014	141	100	100	121	84	85	136	103	97

*For Dataset 3, this table shows a comparison of the three trading strategies annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for $\delta = 0.5\%$ and for gross exposure constraints $c = 5$, 10, and 15. The sample size used was $n = 500$.*

Table 7.15: Balances for Dataset 4 with $c = 5,\ 10,\ 15$.

Year	$c = 5$			$c = 10$			$c = 15$		
	FACE	SAM	FAN	FACE	SAM	FAN	FACE	SAM	FAN
1995	153	100	100	132	121	113	135	108	101
1996	156	100	100	195	136	144	207	171	186
1997	208	100	100	266	183	185	261	199	215
1998	129	96	96	116	99	94	115	92	92
1999	277	130	153	192	155	160	181	130	155
2000	322	149	130	270	163	128	267	174	130
2001	244	172	168	231	150	148	231	143	148
2002	181	167	179	179	149	193	179	148	193
2003	251	140	134	263	157	122	264	153	115
2004	197	94	93	184	103	86	184	125	97
2005	192	97	97	199	121	114	203	134	127
2006	185	100	100	212	111	110	213	135	136
2007	172	100	100	224	181	169	206	111	108
2008	109	100	100	195	97	114	179	101	125
2009	192	136	138	159	72	76	156	68	72
2010	139	134	135	127	113	104	128	105	101
2011	127	87	87	148	97	99	153	116	114
2012	119	111	111	148	76	76	145	100	96
2013	183	72	72	169	175	169	168	135	122
2014	143	85	83	155	75	77	168	105	96

*For Dataset 4, this table shows a comparison of the three trading strategies annual balance (rounded to the nearest pound), assuming an initial balance of £*100 *at the start of each year, for $\delta = 0.5\%$ and for gross exposure constraints $c = 5,\ 10,$ and 15. The sample size used was $n = 500$.*

Figure 45: Sharpe ratios ($c = 5$)

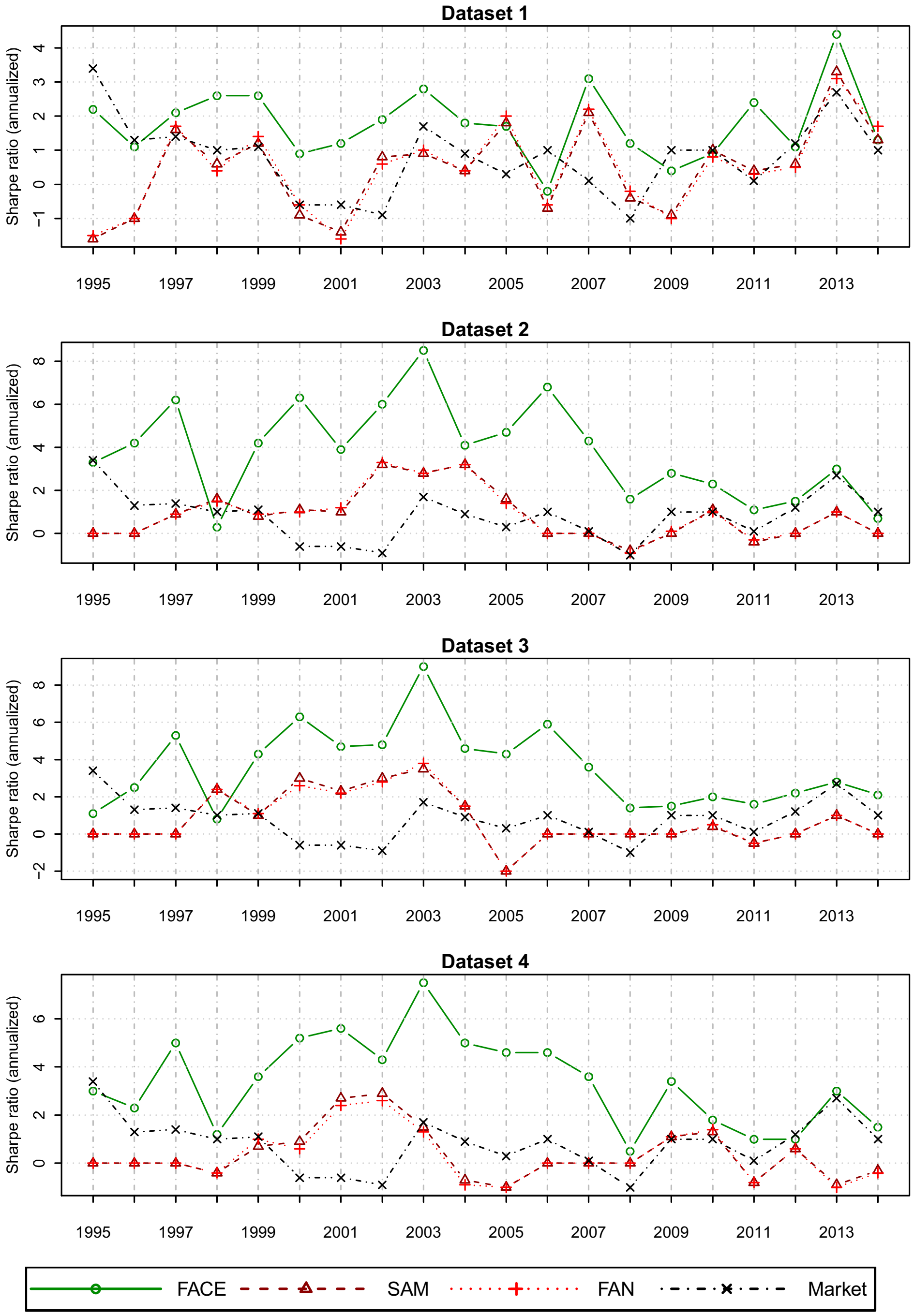


Figure 46: Sharpe ratios ($c = 10$)

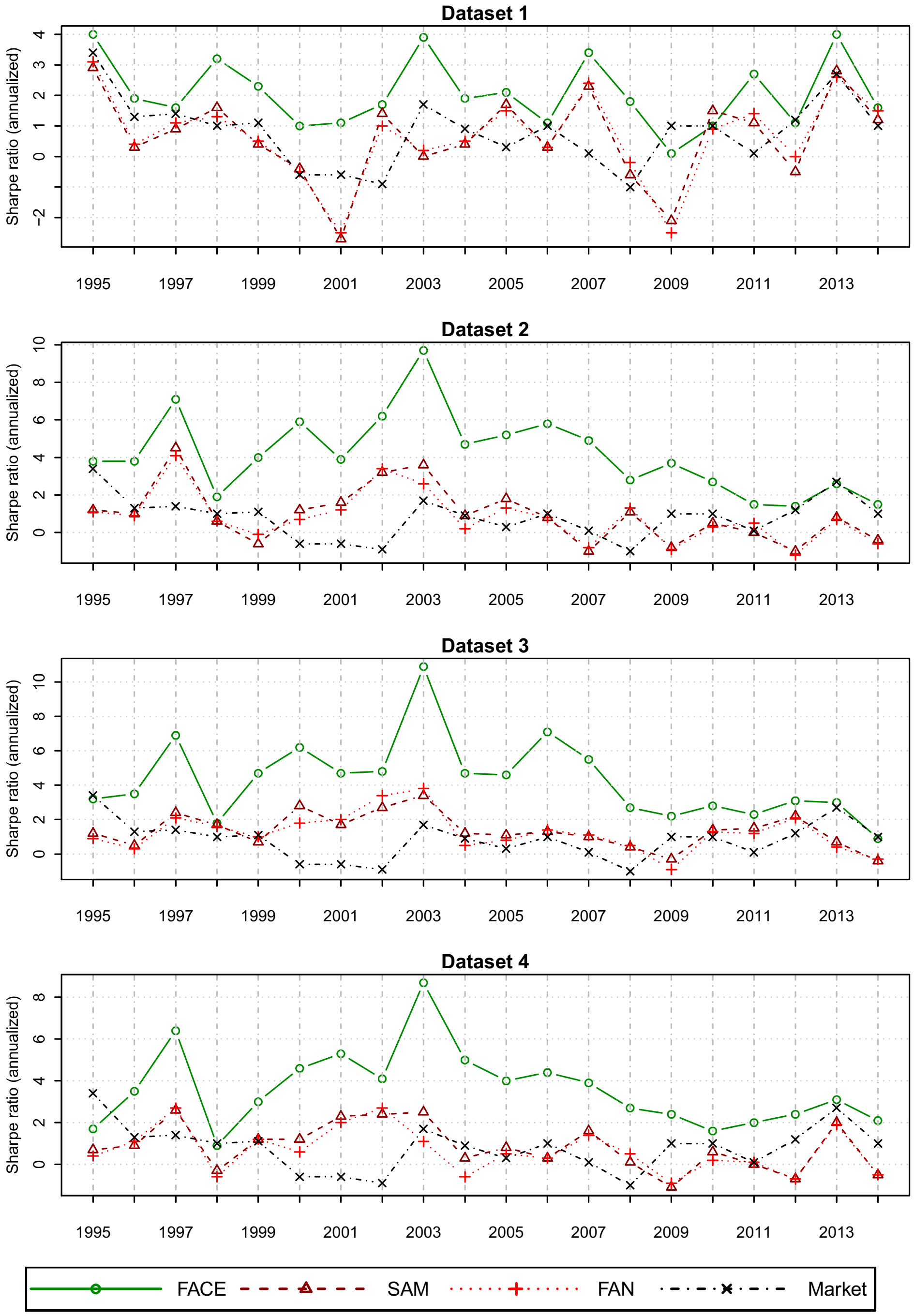


Figure 47: Sharpe ratios ($c = 15$)

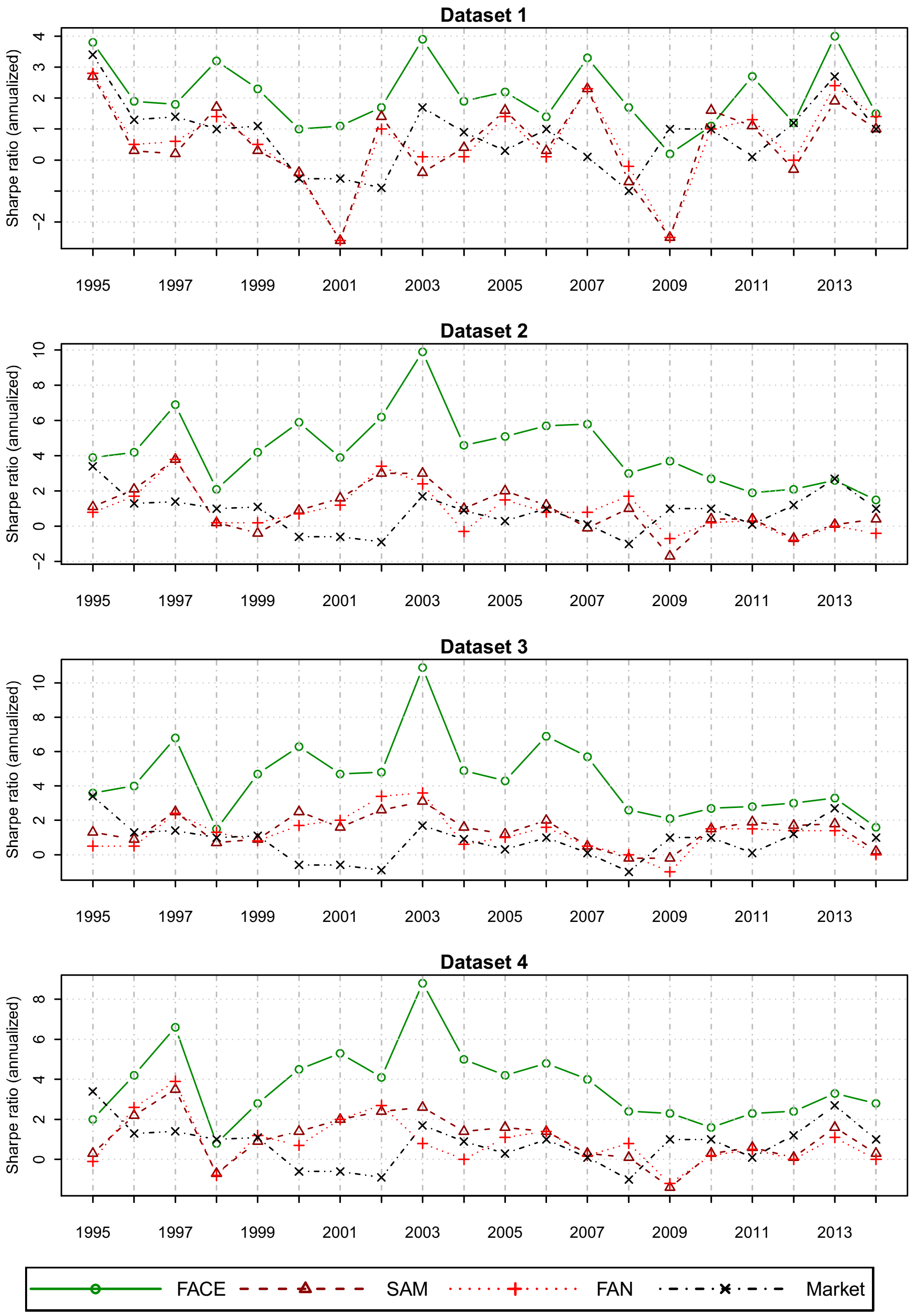
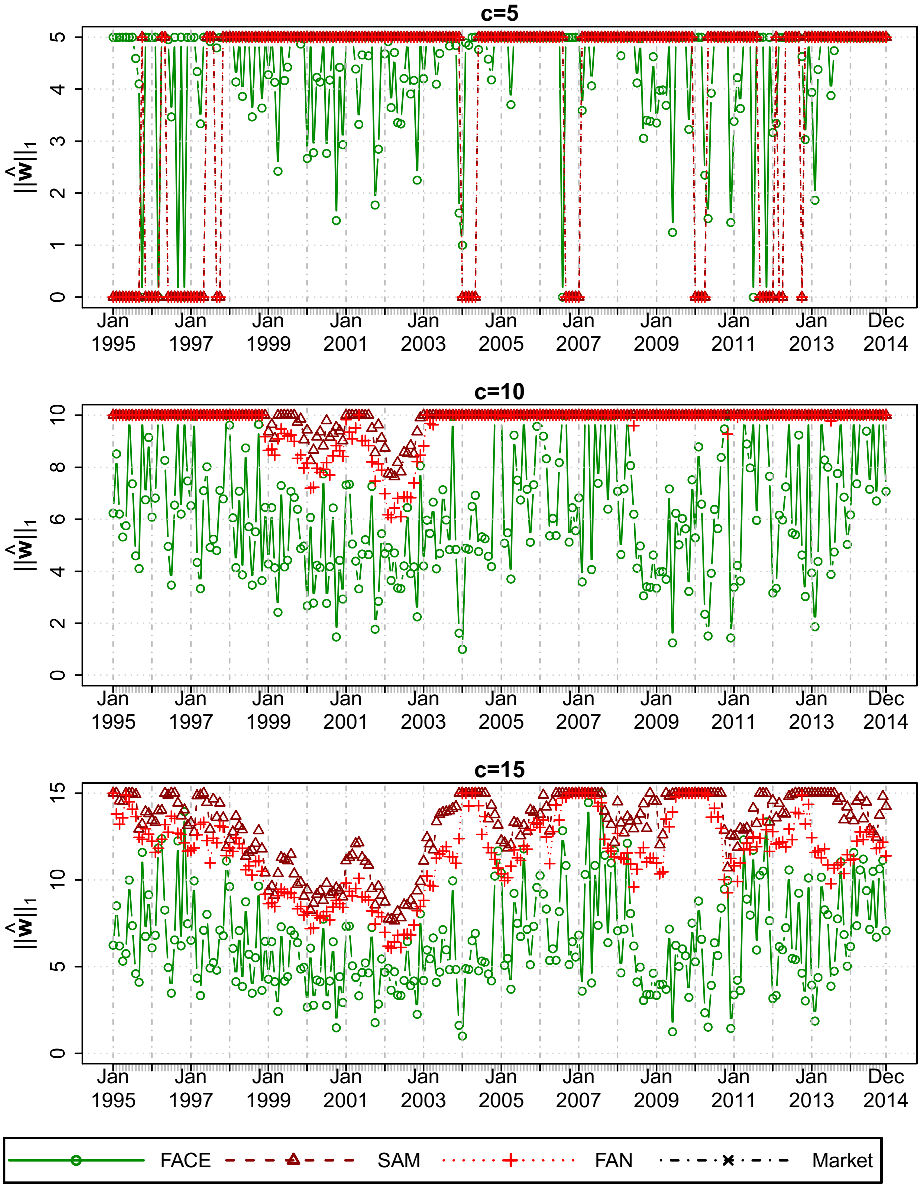
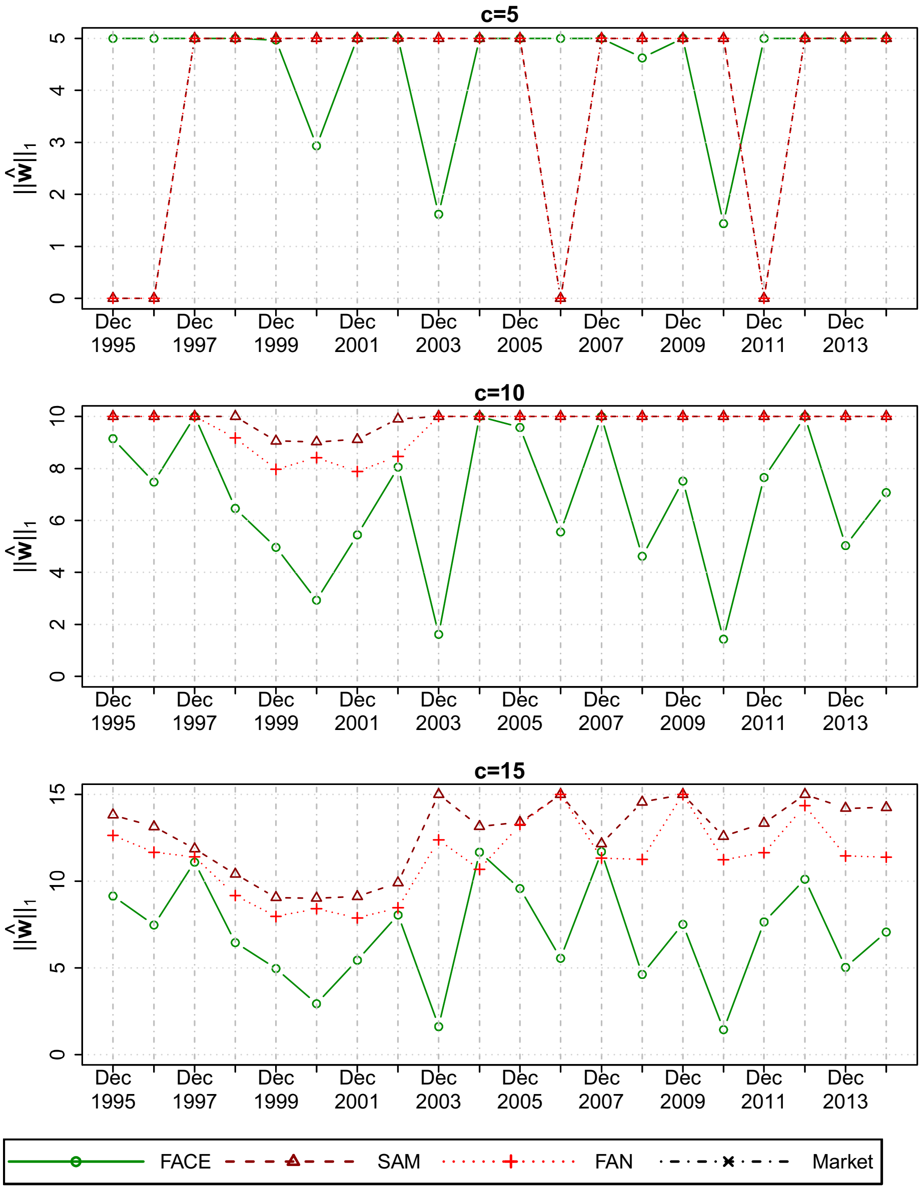


Figure 48: End-of-month gross exposures



This figure shows the end-of-month gross exposures $\|\mathbf{w}\|_1$ *of the three portfolio allocations* $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$ *for* $c = 5$, 10 *and* 15 *in the data analysis for Dataset 1 in Section 7.6.*

Figure 49: End-of-year gross exposures



This figure shows the end-of-year gross exposures $\|\mathbf{w}\|_1$ *of the three portfolio allocations* $\hat{\mathbf{w}}_{\text{FACE}}$, $\hat{\mathbf{w}}_{\text{SAM}}$ *and* $\hat{\mathbf{w}}_{\text{FAN}}$ *for* $c = 5$, 10 *and* 15 *in the data analysis for Dataset 1 in Section 7.6.*

8 Modification of index

In this chapter we explore one possible way to generalise the proposed dynamic structure. Recall that the model introduced in this thesis involves the index $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$. One may argue that this could be improved further if we were to consider more than just the previous trading day inside the index. We explore a slight modification which accommodates for this without increasing the number of unknown parameters of $\boldsymbol{\beta}$. We show, however, that the originally proposed methodology, which only uses the previous trading day, works best when applied to real data.

8.1 Model specification

Our model structure can be trivially extended to accommodate more trading days at a cost of more unknown parameters to estimate. That is, if an extra d trading days were to be included, the number of unknown parameters in $\boldsymbol{\beta}$ would increase from q to $q \times d$. However, this may potentially hinder the estimation performance due to an increase of variance. In order to make our model more flexible, without increasing the number of unknown parameters, one may consider modifying the index using a moving average.

As before, we still assume that $\{(X_t, Y_t),\ t = 1, \cdots, n\}$ is a time series where Y_t denotes a vector of p_n response variables and X_t denotes a vector of q (observable) factors. Further, we still assume that $p_n \longrightarrow \infty$ as $n \longrightarrow \infty$, q is fixed, and that $\{X_t,\ t = 1, \cdots, n\}$ is a stationary Markov process. However, by introducing a simple modification to the index, one may consider

$$Y_t = \mathbf{g}({V_t}^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}({V_t}^{\mathrm{T}}\boldsymbol{\beta})X_t + \boldsymbol{\epsilon}_t, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0 \tag{8.1}$$

where $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$, $\boldsymbol{\Phi}(\cdot)$, $\boldsymbol{\epsilon}_t$ are as before, and $V_t = \frac{1}{\eta}\sum_{\tau=1}^{\eta} X_{t-\tau}$ is a q-dimensional (known) vector depending on some positive integer η. We still assume $\{\boldsymbol{\epsilon}_t,\ t = 1, \cdots, n\}$ is independent of $\{X_t,\ t = 1, \cdots, n\}$, and we further assume that $E(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l : l < t\}) = \mathbf{0}$ and that

$$\mathrm{cov}(\boldsymbol{\epsilon}_t|\{\boldsymbol{\epsilon}_l : l < t\}) = \boldsymbol{\Sigma}_{0,t} = \mathrm{diag}\{\sigma_{1,t}^2, \cdots, \sigma_{p_n,t}^2\}$$

where

$$\sigma_{k,t}^2 = \alpha_{k,0} + \sum_{i=1}^{m} \alpha_{k,i}\epsilon_{k,t-i}^2 + \sum_{j=1}^{s} \gamma_{k,j}\sigma_{k,t-j}^2, \quad t = 2, \cdots, n$$

for each $k = 1, \cdots, p_n$ and for some integers m and s.

Note that (8.1) now contains the index $V_t^{\mathrm{T}}\boldsymbol{\beta}$ as opposed to $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$. This has the new interpretation that the coefficient functions are now depending on the previous η trading days. One can intuitively understand this to mean that the impact of today's factors X_t on today's excess returns Y_t is an unknown function depending on V_t as opposed to X_{t-1}. The original model proposed in this thesis is the special case where $\eta = 1$.

Let $\mathcal{F}_t$ be the σ-algebra generated by $\{(X_l^{\mathrm{T}}, \boldsymbol{\epsilon}_l^{\mathrm{T}}) :\ l \leq t\}$. Assuming for the moment that η is given, and that V_t is known to us at time $t-1$, it is possible to obtain an expression for the conditional covariance matrix $\mathrm{cov}(Y_t|\mathcal{F}_{t-1})$ by taking the conditional covariance of both sides of (8.1), yielding

$$\mathrm{cov}(Y_t|\mathcal{F}_{t-1}) = \boldsymbol{\Phi}(V_t^{\mathrm{T}}\boldsymbol{\beta})\boldsymbol{\Sigma}_x(X_{t-1})\boldsymbol{\Phi}(V_t^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}} + \boldsymbol{\Sigma}_{0,t}$$

where $\boldsymbol{\Sigma}_x(X_{t-1}) \equiv \mathrm{cov}(X_t|X_{t-1})$. In a similar way, by taking conditional expectations, we have

$$E(Y_t|\mathcal{F}_{t-1}) = \mathbf{g}(V_t^{\mathrm{T}}\boldsymbol{\beta}) + \boldsymbol{\Phi}(V_t^{\mathrm{T}}\boldsymbol{\beta})E(X_t|X_{t-1}). \tag{8.2}$$

As with Section 6.2, the goal is to obtain the substitution estimators

$$\widehat{\mathrm{cov}}(Y_t|\mathcal{F}_{t-1}) = \hat{\boldsymbol{\Phi}}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\Sigma}}_x(X_{t-1})\hat{\boldsymbol{\Phi}}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}})^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}_{0,t} \tag{8.3}$$

and

$$\hat{E}(Y_t|\mathcal{F}_{t-1}) = \hat{\mathbf{g}}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\Phi}}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}})\hat{E}(X_t|X_{t-1}). \tag{8.4}$$

Assuming for the moment that η is known, we can estimate $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$, $\boldsymbol{\Phi}(\cdot)$, $\boldsymbol{\Sigma}_x(\cdot)$, and $\boldsymbol{\Sigma}_{0,t}$ using identical methodology to that proposed in Section 6.2 simply by replacing $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$ with $V_t^{\mathrm{T}}\boldsymbol{\beta}$ inside the index. In light of the results from previous chapters we estimate $\boldsymbol{\beta}$ using $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and h_2 using $h_{\hat{k}_{\mathrm{CV}}}$.

One can consider fixing η like we did in the previous chapters of this thesis. Indeed, we shall see that $\eta = 1$, the original model explored in this thesis, performs significantly better when applied to real data. However, one may argue that η can be chosen using a data driven approach using a similar cross validation statistic

$$\sum_{t=n-\nu}^{n} \left\| Y_t - \hat{\mathbf{g}}^{(t-1)}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\Phi}}^{(t-1)}(V_t^{\mathrm{T}}\hat{\boldsymbol{\beta}})X_t \right\| \tag{8.5}$$

where $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are the respective estimates of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ based on $(X_l^{\mathrm{T}}, Y_l^{\mathrm{T}})$, $l = 1, \cdots, t-1$, and where ν is a look-back integer such that $\nu < n-1$. We denote (8.5) by $\mathrm{CV}(k,\eta)$ when $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are estimated with a nearest neighbour bandwidth $h_k(\cdot)$

and where $V_t = \frac{1}{\eta}\sum_{\tau=1}^{\eta} X_{t-\tau}$. We choose k and η by

$$(\hat{k}_{\mathrm{CV}}, \hat{\eta}_{\mathrm{CV}}) = \underset{k,\eta}{\operatorname{argmin}}\big\{\mathrm{CV}(k,\eta)\big\}. \tag{8.6}$$

In order to implement this minimisation we employ a two dimensional grid search algorithm:

(Step 1) Choose a number of candidate nearest neighbour bandwidths, for example: $k \in \mathcal{K}$ where $\mathcal{K} = \{1, \cdots, n\}$.

(Step 2) Choose a number of candidate values of η, for example: $\eta \in \mathcal{A}$ where $\mathcal{A} = \{1, \cdots, A_0\}$ for some integer A_0.

(Step 3) Compute $\mathrm{CV}(k,\eta)$ for each $k \in \mathcal{K}$ and $\eta \in \mathcal{A}$. Choose $(\hat{k}, \hat{\eta}) = \underset{k\in\mathcal{K},\eta\in\mathcal{A}}{\operatorname{argmin}}\big\{\mathrm{CV}(k,\eta)\big\}$.

The above grid search approach has the drawback of an increase in computational time compared to the one dimensional grid search in Section 5.2.

8.2 Comparison with the original approach

In this section, we examine whether one can improve upon $\hat{\mathbf{w}}_{\mathrm{FACE}}$ by choosing a portfolio allocation using the model structure introduced in Section 8.1. We denote by $\hat{\mathbf{w}}(\eta)$ the portfolio allocation using Markowitz's formula, described in Section 6.3, where (8.3) and (8.4) use parameter η in V_t. Hence, the special case $\hat{\mathbf{w}}(1)$ corresponds to $\hat{\mathbf{w}}_{\mathrm{FACE}}$. We fix $\delta = 1.0\%$ and $n = 500$.

We compare the portfolio allocations, $\hat{\mathbf{w}}(1)$, $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$, year by year from Jan 3rd 1995 to Dec 31st 2014 using the same trading strategy which we used in Section 7.2. We only present the results for Dataset 1 for brevity, but similar conclusions can be made for the other three. For each of the portfolio allocations, the end of year balances are reported in Table 8.1 and plots of the Sharpe ratios and maximum drawdowns in Figure 50. As one would expect, there are some similarities between their performances. But overall, it is clear that the originally proposed model, $\hat{\mathbf{w}}(1)$, outperforms $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$ for the vast majority of years.

We remark that although $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$ do perform quite well in some years, and actually slightly beating $\hat{\mathbf{w}}(1)$ occasionally, there are multiple occurrences of significant losses during the years 2000, 2001, 2006, 2009 and 2012 in Table 8.1. These losses are not as severe as $\hat{\mathbf{w}}_{\mathrm{SAM}}$'s or $\hat{\mathbf{w}}_{\mathrm{FAN}}$'s, however. The crucial point is that all these losses could have been avoided simply by using $\hat{\mathbf{w}}_{\mathrm{FACE}}$ instead. The only year that $\hat{\mathbf{w}}_{\mathrm{FACE}}$ slightly lost money was during 2009, as we remarked upon in Section 7.2, and the remaining 19 years all yielded substantial profits.

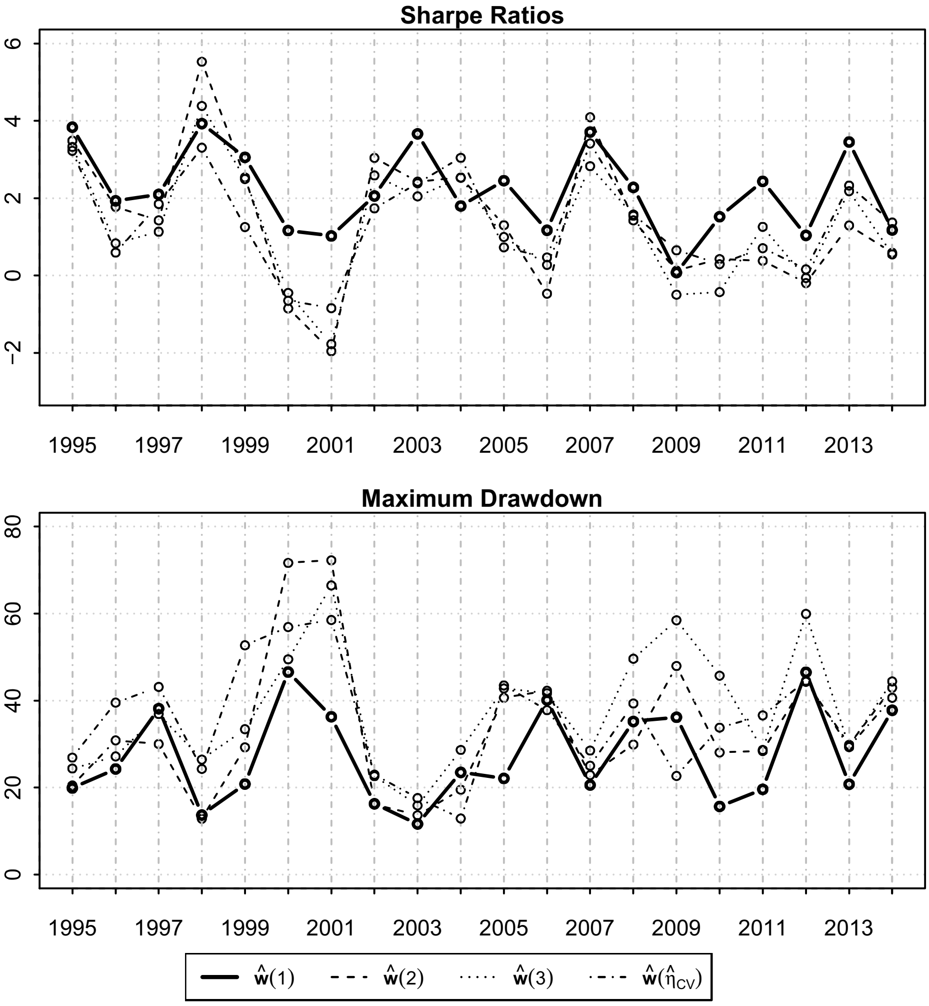
These conclusions can be also seen by the large maximum drawdowns of $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$, implying that these allocations carry significantly larger risk than $\hat{\mathbf{w}}(1)$. Further to this, there are multiple instances of low and negative Sharpe ratios of $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$ during 2000, 2001, 2006, 2009, 2010 and 2012. Also, for the vast majority of time, the Sharpe ratios of $\hat{\mathbf{w}}(1)$ are much larger than the modified index approaches. It is for these reasons, and the conclusions from the real data analysis in Chapter 7, that we recommend only using the originally proposed model $\hat{\mathbf{w}}(1) = \hat{\mathbf{w}}_{\mathrm{FACE}}$.

Table 8.1: End of year balances using modified index

	$\hat{\mathbf{w}}(1)$	$\hat{\mathbf{w}}(2)$	$\hat{\mathbf{w}}(3)$	$\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$	$\hat{\mathbf{w}}_{\mathrm{SAM}}$	$\hat{\mathbf{w}}_{\mathrm{FAN}}$
1995	428	406	419	384	375	475
1996	228	201	133	119	102	121
1997	232	169	151	214	94	125
1998	420	1003	678	368	361	285
1999	332	298	303	166	114	131
2000	163	51	66	60	55	43
2001	140	36	43	62	10	6
2002	206	274	238	177	215	180
2003	273	184	172	184	54	75
2004	179	231	235	256	74	62
2005	265	161	142	126	293	236
2006	151	74	103	112	103	77
2007	563	598	342	445	472	590
2008	356	219	226	247	38	33
2009	94	95	63	121	4	3
2010	153	109	84	104	224	143
2011	284	107	159	125	134	161
2012	148	82	87	97	72	69
2013	393	169	251	250	229	371
2014	166	119	116	182	118	186

Using Dataset 1 from Chapter 7, we apply the identical trading strategy as before but using the portfolio allocations $\hat{\mathbf{w}}(1)$, $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$ which use the modified index approach. The allocation $\hat{\mathbf{w}}(1)$ is equivalent to $\hat{\mathbf{w}}_{\mathrm{FACE}}$. This table shows the balance at the end of the final trading day of each year for each trading strategy.

Figure 50: Comparison of $\hat{\mathbf{w}}(1)$, $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$



Using Dataset 1 from Chapter 7, we apply the identical trading strategy as before but using the portfolio allocations $\hat{\mathbf{w}}(1)$, $\hat{\mathbf{w}}(2)$, $\hat{\mathbf{w}}(3)$, and $\hat{\mathbf{w}}(\hat{\eta}_{\mathrm{CV}})$ which use the modified index approach. The allocation $\hat{\mathbf{w}}(1)$ is equivalent to $\hat{\mathbf{w}}_{\mathrm{FACE}}$. These figures show the Sharpe ratios and the maximum drawdowns for each year.

9 Conclusions and future work

In this chapter we shall outline the main conclusions from this thesis and suggest a possibility for future work using a pursuit of homogeneity.

9.1 Conclusions

In Chapter 3 we explored a simplification of our dynamic structure related to, but slightly different, to an adaptive varying coefficient linear model. As a by product of this thesis, the proposed methodology was shown to have applications outside the field of covariance matrix estimation and portfolio allocation since it can be directly applied to adaptive varying coefficient models. We showed extensive numerical evidence to suggest that our proposed methodology for estimating $\boldsymbol{\beta}$ outperforms the method introduced in Fan et al. (2003).

We introduced a multivariate factor model in Chapter 4 and explored the estimation of $\boldsymbol{\beta}$ using various extensions of the univariate case: $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$. Through simulation studies we found that the performance of $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ and $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ are both significantly better than $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{F}}}$. We also recommended choosing $\hat{\boldsymbol{\beta}}_{\mathrm{L}}$ over $\hat{\boldsymbol{\beta}}_{\bar{\mathrm{L}}}$ since it is slightly more robust to the initial value used in the iterative estimation procedure and faster to compute.

A discussion on bandwidth selection was given in Chapter 5. We saw that the choice of bandwidth is not crucial for estimating $\boldsymbol{\beta}$ as long as it is within a reasonable range. Since bandwidth selection is a computationally expensive task, we recommended only carrying out bandwidth selection on h_2, used in the nonparametric estimation of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$. We explored various data driven bandwidth selectors, implemented using a grid search approach, and found numerical evidence via a simulation study to suggest using cross validation using a nearest neighbour bandwidth.

In Chapter 6 we built upon the work of previous chapters by using the methodology for estimating $\boldsymbol{\beta}$, $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ to estimate the (conditional) covariance matrix. We introduced methodology for estimating the idiosyncratic covariance matrix $\boldsymbol{\Sigma}_{0,t}$ and the conditional covariance matrix $\boldsymbol{\Sigma}_x(\cdot)$. Further, we saw how a simple modification of Markowitz's formula can be used to dynamically allocate a portfolio. Through a simulated example, we saw that $\hat{\boldsymbol{\Sigma}}_{\text{SAM}}$ or $\hat{\boldsymbol{\Sigma}}_{\text{FAN}}$ can suffer from large approximation error since they ignore the dynamic structure. On the other hand, the proposed estimator $\hat{\boldsymbol{\Sigma}}_{\text{FACE}}$ is shown to perform much better, both in terms of the approximation error of estimating the covariance matrix and in terms of the Sharpe ratio related to the estimated portfolio allocation.

A numerical example using real data was presented in Chapter 7. To compare the proposed (dynamic) estimator $\hat{\boldsymbol{\Sigma}}_{\text{FACE}}$ with (constant) estimators $\hat{\boldsymbol{\Sigma}}_{\text{SAM}}$ or $\hat{\boldsymbol{\Sigma}}_{\text{FAN}}$, we looked at a simple trading strategy based on Markowitz's formula. We separately analysed four real datasets, each consisting of approximately 5000 trading days worth of daily returns, and showed that performance of portfolio allocations resulting from $\hat{\boldsymbol{\Sigma}}_{\text{FACE}}$ significantly outperforms both $\hat{\boldsymbol{\Sigma}}_{\text{SAM}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{FAN}}$. This can be seen both in terms of the balance of the trading strategy and the Sharpe ratio of returns. We saw evidence to suggest that these same conclusions hold true for a variety of sample sizes n. This suggests that even if $\hat{\boldsymbol{\Sigma}}_{\text{SAM}}$ and $\hat{\boldsymbol{\Sigma}}_{\text{FAN}}$ only used the observations in a carefully chosen moving window, $\hat{\boldsymbol{\Sigma}}_{\text{FACE}}$ still outperforms them. Finally, we imposed an additional constraint to Markowitz's optimisation problem relating to gross exposure constraints, and showed that the Sharpe ratios of $\hat{\boldsymbol{\Sigma}}_{\text{FACE}}$ still dominates $\hat{\boldsymbol{\Sigma}}_{\text{SAM}}$ or $\hat{\boldsymbol{\Sigma}}_{\text{FAN}}$. Indeed, there are many other constraints and similar optimisation problems investors may use to choose a portfolio allocation. However finding a good estimator of the covariance matrix is crucial, and $\hat{\mathbf{\Sigma}}_{\text{FACE}}$ can equally be applied to these problems in a similar way.

A generalisation of the proposed model was introduced in Chapter 8 regarding a modification of index $X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}$. A natural question is whether an improvement could be made to $\hat{\mathbf{\Sigma}}_{\text{FACE}}$ if we were to consider more than the previous trading day inside this index by using a moving average. We showed, however, that it is optimal to simply use the previous trading day when applied to real data. This provided additional evidence for using the originally proposed estimator $\hat{\mathbf{\Sigma}}_{\text{FACE}}$.

9.2 Homogeneity pursuit

Throughout the thesis, we assumed a model structure of the form

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}) + \epsilon_{k,t}, \quad \|\boldsymbol{\beta}\| = 1, \quad \beta_1 > 0. \tag{9.1}$$

There are many ways in which this could possibly be extended, however we offer one suggestion that could be of particular interest in future work. In order to make (9.1) more flexible, one may naively consider assuming

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + \epsilon_{k,t}, \quad \|\boldsymbol{\beta}_k\| = 1, \quad \beta_{k,1} > 0 \tag{9.2}$$

where $\boldsymbol{\beta}_k \equiv (\beta_{k,1}, \cdots, \beta_{k,q})^{\mathrm{T}}$ for $k = 1, \cdots, p_n$. In (9.2), estimation of each $\boldsymbol{\beta}_k$ can be achieved using univariate methodology introduced in Chapter 3. On the one hand, this is advantageous because we no longer assume every asset shares the same $\boldsymbol{\beta}$, which may not be true in reality. On the other hand, due to an increase in the number of unknown parameters to estimate, the estimation may suffer from a significant increase in variance. In reality it is likely that several assets, such as those from the same industrial sectors, will share the same $\boldsymbol{\beta}$. By taking a pursuit of homogeneity, one can assume

$$y_{k,t} = g_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + X_t^{\mathrm{T}}\mathbf{a}_k(X_{t-1}^{\mathrm{T}}\boldsymbol{\beta}_k) + \epsilon_{k,t}, \tag{9.3}$$

$$\|\boldsymbol{\beta}_k\| = 1, \quad \beta_{k,1} > 0, \quad \boldsymbol{\beta}_k = \boldsymbol{\beta}_{B_j} \quad \text{for all } k \in B_j$$

where it is assumed a partition of $\{1, \cdots, p_n\}$, denoted as $\mathcal{B} = (B_1, \cdots, B_K)$ exists, and that $\boldsymbol{\beta}_{B_j}$ is the common vector shared by all indices in B_j. We assume that K is some integer between 1 and p_n, with the (extreme) special cases $K = 1$ and $K = p_n$ corresponding to (9.1) and (9.2) respectively.

If one has additional (prior) information about each asset, such as their industrial sectors, it is possible to choose $\mathcal{B}$ manually. Alternatively, the partition $\mathcal{B}$, and its size K, can also be estimated using a data driven approach. Once $\mathcal{B}$ and its size K has been determined, $\boldsymbol{\beta}_{B_j}$ $j = 1, \cdots, K$ can be estimated using methodology discussed previously. Since it would be computationally intractable to compute all 2^{p_n} partitions, we suggest a more efficient approach for estimating $\mathcal{B}$ and K based on the K-means algorithm.

Temporarily assume that K is known. First, by using (9.2), we estimate $\{\check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}\}$ by employing univariate methodology from Chapter 3. We wish to assign each univariate estimate $\{\check{\boldsymbol{\beta}}_k;\ k = 1, \cdots, p_n\}$ to one of K cluster centroids $\{\hat{\boldsymbol{\beta}}_{\hat{B}_j};\ j = 1, \cdots, K\}$ where $\hat{\mathcal{B}} = (\hat{B}_1, \cdots, \hat{B}_K)$ is an estimated partition of $\{1, \cdots, p_n\}$. For initialisation purposes, let $\tilde{\mathcal{B}} = (\tilde{B}_1, \cdots, \tilde{B}_K)$ be random partition of $\{1, \cdots, p_n\}$ which uses a random seed s. To find $\hat{\mathcal{B}}$, given some K and s, repeat the following two steps until convergence.

Step 1 (assignment) If this is the first iteration, set $\hat{\mathcal{B}}$ equal to $\tilde{\mathcal{B}}$ and then choose $\hat{\boldsymbol{\beta}}_{\hat{B}_j}$ for $j = 1, \cdots, K$ by sampling from $\{\check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}\}$ without replacement. Otherwise, set $\hat{\mathcal{B}}$ and $\hat{\boldsymbol{\beta}}_{\hat{B}_j}$ equal to the values from the previous iteration. Then:

- For each $k \in \{1, \cdots, p_n\}$ and $j \in \{1, \cdots, K\}$, calculate the angle $\theta_{j,k} = \arccos(\hat{\boldsymbol{\beta}}_{\hat{B}_j}^{\mathrm{T}} \check{\boldsymbol{\beta}}_k)$ between $\check{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\beta}}_{\hat{B}_j}$.
- For each $k \in \{1, \cdots, p_n\}$, assign $\check{\boldsymbol{\beta}}_k$ to its closest cluster centroid using the mapping $\psi(k) = \underset{j \in \{1, \cdots, K\}}{\operatorname{argmin}} \{\theta_{j,k}\}$. This gives us the new partition $\hat{\mathcal{B}} = (\hat{B}_1, \cdots, \hat{B}_K)$ where $\hat{B}_j = \{k : \psi(k) = j\}$.

Step 2 (update) For each $j \in \{1, \cdots, K\}$ update cluster centroid $\hat{\boldsymbol{\beta}}_{\hat{B}_j}$ using

$$\hat{\boldsymbol{\beta}}_{\hat{B}_j} = \begin{cases} \dfrac{\sum_{k=1}^{p_n} \check{\boldsymbol{\beta}}_k I(\psi(k), j)}{\sum_{k=1}^{p_n} I(\psi(k), j)} & \text{if } \sum_{k=1}^{p_n} I(\psi(k), j) > 0 \\ \text{re-sample from } \{\check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}\} & \text{if } \sum_{k=1}^{p_n} I(\psi(k), j) = 0 \end{cases}$$

where

$$I(\psi(k), j) = \begin{cases} 1 & \text{if } \psi(k) = j \\ 0 & \text{if } \psi(k) \neq j. \end{cases}$$

Then, re-standardise $\hat{\boldsymbol{\beta}}_{\hat{B}_j}$ so that $\left\|\hat{\boldsymbol{\beta}}_{\hat{B}_j}\right\| = 1$ and that the first component is greater than zero.

We write $\phi(K, s, \check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n})$ to denote the above K-means algorithm to estimate a partition $\hat{\mathcal{B}} = (\hat{B}_1, \cdots, \hat{B}_K)$ using the (preliminary) univariate estimates $\check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}$ given an integer K and a random seed s.

Using the above algorithm, we now briefly discuss how one may choose to select K. Choosing a large K may result in an estimator with low bias and high variance because the number of unknown parameters could be too large. Conversely, choosing a K which is too small will result in smaller variance but possibly larger bias. We need a way to select an optimum K with regards to this trade off. If one has prior information about the number of sectors / industries, one may wish to choose K manually. Instead, we propose a data driven approach below.

We define the following cross validation statistic

$$\sum_{t=n-\nu}^{n} \left\| Y_t - \hat{\mathbf{g}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k) - \hat{\boldsymbol{\Phi}}^{(t-1)}(X_{t-1}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k)X_t \right\| \tag{9.4}$$

where: $\hat{\mathbf{g}}^{(t-1)}(\cdot)$ and $\hat{\boldsymbol{\Phi}}^{(t-1)}(\cdot)$ are the respective estimates of $\mathbf{g}(\cdot)$ and $\boldsymbol{\Phi}(\cdot)$ based on $(X_l^{\mathrm{T}}, Y_l^{\mathrm{T}})$, $l = 1, \cdots, t-1$; ν is some look-back integer;

$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{\hat{B}_j} \quad \text{for all } k \in \hat{B}_j;$$

$$(\hat{B}_1, \cdots, \hat{B}_K) = \phi(K, s, \check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n});$$

and $\check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}$ are (preliminary) univariate estimates resulting from (9.2) based on $(X_l^{\mathrm{T}}, Y_l^{\mathrm{T}})$, $l = 1, \cdots, t-1$. We denote (9.4) by $\mathrm{CV}(K, s)$.

Using the above notation, we propose the following algorithm to choose K and estimate the partition $\hat{\mathcal{B}}$.

Step 1 (estimate K) Using a grid search approach, calculate

$$(\hat{K}, s) = \underset{K \in \mathcal{K}, s \in \mathcal{S}}{\operatorname{argmin}} \; \mathrm{CV}(K, s)$$

where $\mathcal{K} = \{1, \cdots, p_n\}$, $\mathcal{S} = \{s_1, \cdots, s_r\}$, and $s_1, \cdots, s_r$ are r seeds which are used in the random initialisation of the K-means algorithm. This can be important because the K-means algorithm may finish at a local optimum if a bad initialisation was chosen.

Step 2 (estimate partition $\mathcal{B}$) Using the selected $\hat{K}$ from Step 1, one can estimate the partition $\mathcal{B}$ by running the K-means algorithm multiple times, using different random seeds, to minimise the residual sum of squares. That is, choose

$$\hat{\boldsymbol{\mathcal{B}}} = \underset{s \in \mathcal{S}}{\operatorname{argmin}} \, \mathrm{RSS}(\hat{\boldsymbol{\mathcal{B}}})$$

where

$$\mathrm{RSS}(\hat{\boldsymbol{\mathcal{B}}}) = \sum_{t=2}^{n} \left\| Y_t - \hat{\mathbf{g}}(X_{t-1}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_k) - \hat{\boldsymbol{\Phi}}(X_{t-1}^{\mathrm{T}} \hat{\boldsymbol{\beta}}_k) X_t \right\|^2$$

$$\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_{\hat{B}_j} \quad \text{for all } k \in \hat{B}_j$$

$$\hat{\boldsymbol{\mathcal{B}}} \equiv (\hat{B}_1, \cdots, \hat{B}_{\hat{K}}) = \phi(K, s, \check{\boldsymbol{\beta}}_1, \cdots, \check{\boldsymbol{\beta}}_{p_n}).$$

This is just one approach for estimating $\hat{\boldsymbol{\mathcal{B}}}$ and K, and has not been tested on real data yet. The purpose of presenting it is to show just one example of how the dynamic structure can be potentially improved, and to demonstrate that there is scope for exciting future work.

References

Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, pages 2577–2604.

Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, 2nd edition.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489.

Cheng, M.-Y., Zhang, W., and Chen, L.-H. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, 104(487).

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, pages 309–376.

Eaton, M. L. and Tyler, D. (1994). The asymptotic distribution of singular-values with applications to canonical correlations and correspondence analysis. *Journal of Multivariate Analysis*, 50(2):238–264.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *the Journal of Finance*, 47(2):427–465.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004.

Fan, J., Fan, Y., and Lv, J. (2008a). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186 – 197. Econometric modelling in finance and risk management: An overview.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66 (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).* Chapman & Hall, 1 edition.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Fan, J., Yao, Q., and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):57–80.

Fan, J., Zhang, J., and Yu, K. (2008b). Asset allocation and risk assessment with gross exposure constraints for vast portfolios. *Available at SSRN 1307423*.

Fan, J., Zhang, J., and Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606.

Fan, J. and Zhang, W. (1999). Statistical Estimation in Varying Coefficient Models. *The Annals of Statistics*, 27(5):1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, 27(4):715–731.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185.

Goldfarb, D. and Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis*, pages 226–239. Springer.

Goldfarb, D. and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical programming*, 27(1):1–33.

Green, P. and Silverman, B. (1993). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Hardle, W., Hall, P., Ichimura, H., et al. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

J Fan, T Gasser, I. G. (1995). On nonparametric estimation via local polynomial regression (1995). *Institute of statistics, Catholic University of Louvain, Louvain-la-Neuve,Belgium.*

Jagannathan, R. and Ma, T. (2002). Risk reduction in large portfolios: Why imposing the wrong constraints helps. Technical report, National Bureau of Economic Research.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Karoui, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *The Annals of Statistics*, pages 2717–2756.

Kong, E., Xia, Y., et al. (2014). An adaptive composite quantile approach to dimension reduction. *The Annals of Statistics*, 42(4):1657–1688.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Li, J. and Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association*, 106(494):685–696.

Markowitz, H. (1952). Portfolio selection*. *The journal of finance*, 7(1):77–91.

Markowitz, H. M. (1968). *Portfolio selection: efficient diversification of investments*, volume 16. Yale university press.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22:1346–1370.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):pp. 1–52.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Sun, Y., Yan, H., Zhang, W., Lu, Z., et al. (2014). A semiparametric spatial dynamic model. *The Annals of Statistics*, 42(2):700–727.

Sun, Y., Zhang, W., and Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics*, pages 2795–2814.

Turlach, B. and Weingessel, A. (2013). Functions to solve Quadratic Programming Problems. [Online] Available from: `https://cran.r-project.org/web/packages/quadprog/`. [Accessed: 2nd April 2015].

Wand, P. and Jones, C. (1994). *Kernel Smoothing*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser.*, 26:359–372.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

Wuertz, D. and Chalabi, Y. (2013). Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. [Online] Available from: `https://cran.r-project.org/web/packages/fGarch`. [Accessed: 2nd April 2015].

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162–1184.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054.

Zhang, W., Fan, J., and Sun, Y. (2009). A semiparametric model for cluster data. *The Annals of Statistics*, 37(5A):2377.