Maximum Rank Correlation Estimation for Generalized Varying-Coefficient Models with Unknown Monotonic Link Function

Xiang Li

PhD

University of York

Mathematics

July 2016

Abstract

Generalized varying coefficient models (GVCMs) form a family of statistical utilities that are applicable to real world questions for exploring associations between covariates and response variables. Researchers frequently fit GVCMs with particular link transformation functions. It is vital to recognize that to invest a model with a wrong link could provide extremely misleading knowledge. This thesis intends to bypass the actual form of the link function and explore a set of GVCMs whose link functions are monotonic. With the monotonicity being secured, this thesis endeavours to make use of the maximum rank correlation idea and proposes a maximum rank correlation estimation (MRCE) method for GVCMs. In addition to the introduction of MRCE, this thesis further extends the consideration to Generalized Semi-Varying Coefficient Models (GSVCMs), Panel data, simulations and empirical studies.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, may I sincerely express my appreciation to Prof. Wenyang Zhang, for his guidance and support throughout this project. He is an enthusiastic and deep-minded statistician during working time, and a wise and approachable friend privately. His training and enlightenment will always be valuable to me.

Secondly, I would like to thank all the staff at the Department of Mathematics at the University of York, in particular my Thesis Advisory Panel members - Dr. Marina Knight, Dr. Degui Li and Dr. Samer Kharroubi, for their continuous help and support.

I would give thanks to my Ph.D friends, Dr. Hongjia Yan, Dr. John Box and Dr. Yuan Ke. They make my Ph.D days an interesting and memorable experience that I will treasure in my whole life.

Finally, I would like to thank my father Anwu Li and my mother Jiaju Qin for their continuous support. I want to give special thanks to my wife Lin Chen. Her love and understanding direct me to better myself and become a responsible person.

Author's Declaration

Chapter 2, Literature Review, summarizes some key ideas and precious researchers related to this thesis. In particular:

- Chapter 2.1 reviews the fundamental concepts and ideas of local polynomial modelling. The content summarizes the framework of the first three chapters of the book *Local Polynomial modelling and Its applications.* by Fan and Gijbels (1996).
- Chapter 2.2 is a review of the varying coefficient models, which is mainly based on research works of Hoover *et al*. (1998) and Fan and Zhang (1999,2008).
- Chapter 2.3 summarizes the generalized varying coefficient models introduced by McCullagh and Nelder (1989) and reviews the Maximum Likelihood Estimation method proposed by Cai, Fan and Li (2000).

The remaining chapters are original work containing the Maximum Rank Correlation Estimation method for generalized varying coefficient models with unknown monotonic link function.

To the best of my knowledge and belief the work does not infringe the copyright of any other person. I hereby declare that this thesis is an original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

1 Introduction

The varying coefficient models proposed by Cleveland, Grosse and Shyu (1991) and Hastie and Tibshirani (1993) are useful extensions of linear models. The feature of standard varying coefficient models is to allow the coefficients to be smoothing non-parametric functions, which can be used to explore the dynamics of the impacts of the covariates on the response variable. Due to the generality of the functional coefficient, modelling bias can be reduced significantly and the 'curse of dimensionality' can be avoided (Zhang, Li and Song, 2002). There are wide applications of varying coefficient models in various disciplines. For example, Hoover *et al.*'s (1998) application of the model to longitudinal data; non-linear time series applications of Chen and Tsay (1993) and Cai, Fan and Yao (2000). Varying coefficient models form a very useful framework. There are as well a lot of extensions of varying coefficient models in existing literature, including for instance semi-parametric varying-coefficient partially linear models (Fan and Huang, 2005) and semi-varying coefficient models (Li and Liang, 2008).

Despite the flexibility and interpretability of the varying coefficient models, there are situations where further extensions are preferred. For example, the range of the response variable could be restricted; the variance of the response variable could depend on its mean. These two issues are well addressed by generalized linear models (GLMs) in extension to traditional linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). For details of GLM, see Dobson (1990). There are a flexible class of applications arising from GLMs. However, these parametric applications are not flexible enough to capture the true underlying relationship between covariates and responses (Lian, 2012). It is not always reasonable to suppose that the impact on the response variable is constant. Instead, with respect to certain covariates, the association to the response variable can be varying and complex (Hastic and Tibshirani, 1993). When varying impact exists, it is of interest to expand the application of generalized linear model by relaxing some of its conditions.

The ideas of generalized linear models (GLMs) and varying coefficient models (VCMs) are readily combined. The generalized varying coefficient models (GVCMs) are extensions of both varying coefficient models and generalized linear models. One central technique of GVCMs is the application of link function which describes how the mean response variable depends on the linear predictor. Statistic works of GVCMs frequently assume that the transformation link functions are some known functions. For example, Cai, Fan and Li's (2000) consider Poisson regression with log link function. With this given link function, Local Maximum Likelihood Estimation method can be an ideal approach to access the varying coefficients. Although, their attempt is reasonable for count data analysis, using specified link function can only be useful, rather than be true.

Real world data analysis are different from simulations. We don't really know which kind of function the link function should be. Miss-specified models could be extremely biased. A sensible way would be to let the data specify the link function. This thesis attempts to extend GVCMs by loosening the condition on link function. Instead of assuming that the link transformation follows some specific functional form, the thesis supposes that the functional form is unknown but structural. Our operation is intuitive and practically interesting. In many situations - although one does not in advance acquire the functional form of the link function - it is reasonable to assume that the unknown link function possesses some certain features. In this thesis, a particular group of structural link transformations, the strictly monotonic transformations, are considered. A monotonic transformation preserves a useful probability feature, which the thesis is going to introduce in later context. For this type of models, this thesis utilizes the feature of monotonic link function and estimate the varying coefficients with maximum rank correlation estimation method. Based on the estimators of the varying coefficients, the unknown monotonic function is approached with maximum likelihood estimation.

2 Literature Review

2.1 Local polynomial modelling

Non-parametric modelling methods are powerful in the sense that they allow researchers to relax the assumptions on the forms of regression functions, and let the available data search for the fine structural relationships. It has been demonstrated that the local polynomial modelling methods have vast quality statistical properties and outstanding practise performance. Therefore, the local polynomial methods are treated as very strong tools in non-parametric analysis. Their advantages can be revealed through applications, such as, survival analysis, generalized linear models and time series. In this paper, the understanding of the local polynomial modelling methods and their application is crucial. In the beginning of the thesis, the author briefly reviews the local polynomial modelling discussed in detail by Fan and Gijbels in their 1996 book, *Local Polynomial Modeling and Its Applications*.

The thesis starts by looking at the case of bivariate data. Suppose $(X_i, Y_i), i = 1, \cdots, n$, are *i.i.d*, and generated from the population (X, Y), where X and Y satisfy the model

$$Y = m(X) + \sigma(X)\epsilon. \tag{2.1}$$

In model (2.1), variables X and ϵ are independent, and ϵ is from standard Gaussian distribution. $m(x) = E(Y|X = x)$ is the conditional mean at $X = x$, and the conditional variance is denoted by $\sigma(x) = Var(Y|X = x)$. Further, denote by $f(\cdot)$ the marginal density of X. One is interested in the regression function $m(\cdot)$, and its derivatives $m^{(j)}(\cdot), j = 1, \cdots, p$, where j is referred to as the order of the derivatives.

Instead of fitting the unknown regression function globally, the thesis searches for the regression function values at grid points locally, using only the data in the neighbourhood of each grid point, by Taylors expansion on the base of Least Squares methods. Recall and apply the Taylors expansion of order p for the regression function $m(x)$ in the neighbourhood of x_0,

$$m(x) \approx m(x_0) + m^{(1)}(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \tag{2.2}$$

Denote by

$$\beta_\nu = \frac{m^{(\nu)}(x_0)}{\nu!}, \nu = 0, \cdots, p,$$

where $\beta_0 = m(x_0)$. The approximation (2.2) can be rewrote as

$$m(x) = \sum_{j=0}^{p} \beta_j (x - x_0)^j. \tag{2.3}$$

Suppose X_i, $i = 1, \cdots, n$, are the datum points. In the neighbourhood of x_0, consider the WLS (weighted least squares) problem

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0), \tag{2.4}$$

where $K_h(\cdot) = \frac{K(\cdot/h)}{h}$, $K(\cdot)$ is the kernel function, and h is the controlling bandwidth. Minimization of (2.4) with respect to β_j leads to the estimates of the mean regression function $m(x_0)$ and its derivatives.

With the notations below, the WLS problem (2.4) can be re-expressed in a matrix form, which is more convenient in practice. Given datum points

(X_i, Y_i), $i = 1, \cdots, n$, in the neighbourhood of x_0, denote

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x_0 \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \\ 1 & X_n - x_0 \cdots & (X_n - x_0)^p \end{pmatrix}, \quad y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

$$W = \text{diag}\Big(K_h(X_1 - x_0),\ \cdots,\ K_h(X_n - x_0)\Big),$$

and

$$\beta = \big(\beta_0, \cdots, \beta_p\big)^T.$$

The WLS problem in (2.4) is equivalent to

$$WLS = \min_{\beta}(y - \mathbf{X}\beta)^T W (y - \mathbf{X}\beta). \tag{2.5}$$

Take derivative of (2.5) with respect to β as

$$\frac{\partial WLS}{\partial \beta} = \frac{\partial (y - \mathbf{X}\beta)^T}{\partial \beta} \frac{\partial WLS}{\partial (y - \mathbf{X}\beta)^T} = -\mathbf{X}^T \cdot 2W(y - \mathbf{X}\beta).$$

One can easily obtain the solution vector

$$\hat{\beta} = (\mathbf{X}^T W \mathbf{X})^{-1} \mathbf{X}^T W y, \tag{2.6}$$

where $\hat{\beta} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)^T$ and $\hat{m}_\nu(x_0) = \nu! \hat{\beta}_\nu$, $\nu = 0, \cdots, p$, are the estimates of interest.

To assess the performance of the estimates, a convenient subject to consider is the MSE (Mean Squared Error) or MISE (Mean Integrated Squared

Error). The definitions are as the following:

$$\begin{aligned} MSE(x) &= E\Big[\{\hat{m}_\nu(x) - m^{(\nu)}(x)\}^2|\mathbf{X}\Big] \\ &= \Big[E\{\hat{m}_\nu(x)|\mathbf{X}\} - m^{(\nu)}(x)\Big]^2 + Var\{\hat{m}_\nu(x)|\mathbf{X}\}, \qquad (2.7) \end{aligned}$$

where $\Big[E\{\hat{m}_\nu(x)|\mathbf{X}\}$ - $m^{(\nu)}(x)\Big]^2$ and $Var\{\hat{m}_\nu(x)|\mathbf{X}\}$ are called the conditional bias and variance, respectively;

$$MISE(x) = \int MSE(x)w(x)dx, \qquad (2.8)$$

where $w(\cdot) > 0$ is some weight function.

It is clear that to explore the MSE or the MISE, it is not avoidable to look at the conditional bias and the variance. Derivation of the conditional bias and variance of $\hat{\beta}$ gives

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T W(r + \mathbf{X}\beta) \\ &= \beta + (\mathbf{X}^T W\mathbf{X})^{-1}\mathbf{X}^T W r, \\ \text{and} \quad Var(\hat{\beta}|\mathbf{X}) &= (\mathbf{X}^T W\mathbf{X})^{-1}(\mathbf{X}^T\Sigma\mathbf{X})(\mathbf{X}^T W\mathbf{X})^{-1}, \end{aligned}$$

where $\Sigma = diag\{K_h^2(X_1 - x_0)\sigma^2(X_1), \cdots, K_h^2(X_n - x_0)\sigma^2(X_n)\}$, and $r = m - \mathbf{X}\beta$ is the residual vector.

The conditional bias and variance of $\hat{\beta}$ are not directly usable due to the existence of unknown quantities, r and Σ. One effective way is to approximate the conditional bias and variance by their first order asymptotic expansions. The asymptotic behaviours of the conditional bias and variance are well studied by Ruppert and Wand (1994) and quoted by Fan and Gijbels (1996). Before recapping the asymptotic bias and variance in Theorem 2.1,

a few notations have to be introduced. Let $e_{\nu+1}$ be the indicator vector with the $(\nu+1)_{th}$ entry being set as 1 and other entries being set as 0. Further denote
$\mu_j = \int u^j K(u)du, \quad \nu_j = \int u^j K^2(u)du,$
$S = (\mu_{j+l})_{0\leq j,l\leq p}, \quad \tilde{S} = (\mu_{j+l+1})_{0\leq j,l\leq p},$
$S^* = (\nu_{j+l+1})_{0\leq j,l\leq p}, \quad c_p = (\mu_{p+1}, \cdots, \mu_{2p+1})^T,$ and $\tilde{c}_p = (\mu_{p+2}, \cdots, \mu_{2p+2})^T.$

Theorem 2.1. *Assume that $f(x_0) > 0$ and that $f(\cdot)$, $m^{(p+1)}$ 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*

$$Var(\hat{m}_\nu(x_0)|\mathbb{X}) = e_{\nu+1}^T S^{-1} S^* S^{-1} e_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + op(\frac{1}{nh^{1+2\nu}}); \qquad (2.9)$$

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

$$Bias(\hat{m}_\nu(x_0)|\mathbb{X}) = e_{\nu+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} m^{p+1}(x_0) h^{p+1-\nu} + op(h^{p+1-\nu}); \quad (2.10)$$

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

$$\begin{aligned} Bias(\hat{m}_\nu(x_0)|\mathbb{X}) &= e_{\nu+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+1)!} \Big\{ m^{p+1}(x_0) + (p+2) m^{p+2}(x_0) \\ &\quad \frac{\dot{f}(x_0)}{f(x_0)} h^{p+2-\nu} \Big\} + op(h^{p+2-\nu}), \qquad (2.11) \end{aligned}$$

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

Shown in Theorem 2.1, the performance of local estimates is relevant with many factors, such as, the kernel function, the design density, bandwidth and polynomial order. Intuitively, kernel function determines the way how weights are assigned to the datum points in the neighbourhood; the design density describes the marginal distribution of the datum points; a proper polynomial order selection helps to balance the estimation bias and variance; and the bandwidth selection controls the model complexity. Hence, all of them are important to our local estimation and cannot be chosen arbitrarily. Details of these problems are well discussed by Fan and Gijbels (1996) in *Local Polynomial Modeling and Its Applications*. The thesis does not intend to repeat their work.

2.2 Varying coefficient models

The varying coefficient models (VCMs) are important tools utilized for exploring the dynamic pattern in many scientific areas, such as economics, finance, epidemiology and so on. They are primarily developed from practical needs. Due to the flexibility and interpretability that varying coefficient models possess, there have been solid developments on the models' methodological, theoretical and practical sides (Fan and Zhang, 2008) in the past two decades. In this section, the thesis only briefly recaps the varying coefficient models to the basics.

2.2.1 Model construction

For given scaler U, over which the coefficient functions vary, covariates $X = (x_1, \cdots, x_p)^T$, and response variable y, the varying coefficient model is defined as

$$y = \sum_{j=1}^{p} \beta_j(U) x_j + \epsilon, \tag{2.12}$$

with $E(\epsilon|U, x_1, \cdots, x_p) = 0$, and $Var(\epsilon|U, x_1, \cdots, x_p) = \delta^2(U)$. When $x_1 \equiv 1$, the model permits a varying intercept term. For unknown coefficient functions $\boldsymbol{\beta}(U) = \big(\beta_1(U), \cdots, \beta_p(U)\big)^T$, the multivariate mean regression function is given by

$$m(U, X^T) = \sum_{j=1}^{p} \beta_j(U) x_j,$$

where $E(y|U, X^T) = m(U, X^T)$.

There are a few disciplines that the varying coefficient functions are estimated. For example, local polynomial modelling(Hoover *et al.*, 1998; Fan and Zhang, 1999), polynomial spline (Huang *et al.*, 2002,2004; Huang and

Shen, 2004) and smoothing spline (Hastie and Tibshirani, 1993; Chiang *et al.*, 2001). Since the varying coefficient models are primarily locally linear models, Fan and Zhang (2008) claim that it is more reasonable to apply local polynomial smoothing methods.

2.2.2 Estimation of the coefficient functions

Suppose that $(U_i, X_i^T, y_i), i = 1, \cdots, n$, consists a sample of (U, X^T, y) from model (2.12). For each given u, the local linear estimator $\hat{\boldsymbol{\beta}}(u)$ of $\boldsymbol{\beta}(u)$ is the part corresponding to $\mathbf{a}$ of the minimizer of

$$L(\mathbf{a}, \dot{\mathbf{a}}) \sum_{i=1}^{n} \Big\{ y_i - X_i^T \mathbf{a} - X_i^T \dot{\mathbf{a}}(U_i - u) K_h(U_i - u) \Big\}, \tag{2.13}$$

where h is the smoothing bandwidth and $K_h(\cdot) = \frac{K(\cdot/h)}{h}$. $K(\cdot)$ is the kernel function which is usually taken to be the Epanechnikov kernel $K(t) = 0.75(1-t)_+$. The estimator is normally distributed, asymptotically.

Let $\mathbf{X} = (X_1, \cdots, X_n)^T$, $\mathbf{U} = diag(U_1 - u, \cdots, U_n - u)$, $Y = (y_1, \cdots, y_n)^T$, $W = diag\Big(K_h(U_1 - u), \cdots, K_h(U_n - u)\Big)$, and $\Gamma = \mathbf{X}^T \mathbf{U} \mathbf{X}$. Then the estimator $\hat{\boldsymbol{\beta}}(u)$ is given by

$$\hat{\boldsymbol{\beta}}(u) = (I_p, 0_p)(\Gamma^T W \Gamma)^{-1}(\Gamma^T W Y), \tag{2.14}$$

where I_p is a $p-$dimensional identity matrix and 0_p is a size p matrix with each entry being 0.

2.3 Generalized varying coefficient models

The generalized varying models are widely explored and used in statistical applications. Like the generalized linear models (McCullagh and Nelder, 1989), the generalized varying models provide a framework for relating response and predictor variables. The generalized varying coefficient models allow the regression coefficients to vary depending on certain covariates, for instance, age and time, which widens the scope and applicability in practice. In the work of Cai, Fan and Li (2000), they consider a family of generalized varying coefficient models with given link functions.

2.3.1 Model construction

Suppose U is a scalar over which the coefficient functions vary, $\mathbf{X} = (X_1, \cdots, X_p)^T$ holds the covariates, and y is the response variable. The generalized varying models basically have two assumptions. Firstly, the conditional distribution of y given $\mathbf{X} = \mathbf{x}$ is from the popular exponential family

$$f(y|\mathbf{X} = \mathbf{x}) = exp\left\{\frac{\theta(\mathbf{x})y - b\left(\theta(\mathbf{x})\right)}{a(\phi)} + c(y, \phi)\right\},$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are known functions; $\theta(\cdot)$ and ϕ are the canonical and dispersion parameters, respectively. The exponential family includes many commonly applied distributions, such as Gaussian, Poisson and Gamma distribution.

Secondly, for each U, a generalized varying coefficient model gives

$$g\Big\{m(U, \mathbf{X})\Big\} = \mathbf{X}^T \boldsymbol{\beta}(U), \tag{2.15}$$

where $m(U, \mathbf{X}) = E(y|U, \mathbf{X})$ is the conditional mean regression function,

$\boldsymbol{\beta}(U)$ is the vector of varying coefficients, and $g(\cdot)$ is referred to as the link function.

When the link function is provided, for example a log function or a logit transformation, one is able to access the association between the covariates and the responses by utilizing statistical estimation methods and hypothesis testing techniques. The estimation of the varying coefficients with given link function $g(\cdot)$ with maximum likelihood estimation method is discussed in detail by Cai, Fan and Li (2000). Since the interest is in the accessibility to the association between the covariates and the response variable, the thesis is only to recall the estimation of the varying coefficients with given link transformation.

2.3.2 Estimation of the varying coefficient functions

Local linear fittings are statistically efficient and design-adaptive (Fan, 1993), and have nice boundary performance (Fan and Gijbels, 1996). At any point u, with local linear approximation to the varying coefficients,

$$\boldsymbol{\beta}(U_i) = \boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_i - u)$$

, the estimator $\hat{\boldsymbol{\beta}}(u)$ of $\boldsymbol{\beta}(u)$ is the part corresponding to $\mathbf{a}$ of the maximizer of the local log-likelihood function

$$L(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[g^{-1}\big(\mathbf{X}_i{}^T\mathbf{a} + \mathbf{X}_i{}^T\mathbf{b}(U_i - u)\big),\ y_i \right] K_h(U_i - u), \qquad (2.16)$$

where h is the smoothing bandwidth, $K_h(\cdot) = \frac{K(\cdot/h)}{h}$, and $K(\cdot)$ is the kernel function.

Let $\mathbf{Z}_i = \big(\mathbf{X}_i{}^T, \mathbf{X}_i{}^T(U_i - u)\big)$, and $\mathbf{B} = (\mathbf{a}, \mathbf{b})^T$. To make the presentation clear, use $d(\cdot)$ to represent $g^{-1}(\cdot)$. The maximizer of (2.16) can be approximated by the maximizer of

$$L(\mathbf{B}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left[d\big(\mathbf{Z}_i{}^T\mathbf{B}\big),\ y_i\right] K_h(U_i - u).$$

Its first and second derivatives are given by

$$\dot{L}(\mathbf{B}) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell' \left[d(\mathbf{Z}_i{}^T\mathbf{B}),\ y_i\right] d'(\mathbf{Z}_i{}^T\mathbf{B})\mathbf{Z}_i K_h(U_i - u),$$

and

$$\begin{aligned}\ddot{L}(\mathbf{B}) \;=\; & \frac{1}{n}\sum_{i=1}^{n} \Big\{\ell'' \left[d(\mathbf{Z}_i{}^T\mathbf{B}),\ y_i\right] d'^2(\mathbf{Z}_i{}^T\mathbf{B}) + \ell' \left[d(\mathbf{Z}_i{}^T\mathbf{B}),\ y_i\right] d''(\mathbf{Z}_i{}^T\mathbf{B})\Big\} \\ & \times \mathbf{Z}_i{}^T\mathbf{Z}_i K_h(U_i - u).\end{aligned}$$

The Newton-Raphson maximization algorithm updates the estimator as follows. Let $\mathbf{B}_n$ be current $\mathbf{B}$, $\mathbf{B}$ is then updated via

$$\mathbf{B}_{n+1} = \mathbf{B}_n - \ddot{\ell}(\mathbf{B}_n)^{-1}\dot{\ell}(\mathbf{B}_n)$$

until convergence, which gives the estimates for $\mathbf{B}$. Denote $\hat{\mathbf{B}} = \left(\hat{\mathbf{a}}, \hat{\mathbf{b}}\right)^T$, $\hat{\boldsymbol{\beta}}(u) = \hat{\mathbf{a}}$ is the estimators of the varying coefficients. In practise, one could face singular matrix problem, Cai, Fan and Li (2000) suggest to take into account the idea of ridge regression (Seifert and Gausser, 1996; Fan and Chen, 1999), and use ridge parameters to tackle singular or nearly singular matrix problems.

3 Generalized Varying Coefficient Models with Unknown Monotonic Link Function

In real world applications, it is common that researchers construct generalized varying coefficient models with specific link functions. For example, log transformation is frequently applied for count data, and logit transformation is mostly considered for binary data. However, it is not ideal to 'guess' a function and assume its validity. As a wrong model could be extremely biased, it is more desirable to let the link function be data specified. With emphasis on regression coefficients estimation, generation of more applicable statistical methods are therefore possible. Indeed, we are blinded if the link function is completely unknown. Whereas, by assuming that the link function enjoys some certain properties, it is possible to undermine the hidden structure of the varying coefficients.

3.1 Model Assumption

Let $(\mathbf{X}_i^{\mathrm{T}},\ U_i,\ y_i)$, $i = 1,\ \cdots,\ n$, be an *i.i.d.* sample from some certain population $(\mathbf{X}^{\mathrm{T}},\ U,\ y)$, where y is the response variable, $\mathbf{X}^T = X_1, \cdots, X_p{}^T$ is the corresponding p-dimensional covariates, and U on which the varying coefficients depend is a scalar/index. For simplicity, only univariate U is considered in this thesis. Denote the mean regression function of y give $\mathbf{X}^T$ and U by $m(\mathbf{X}^T, U) = E(y|\ U,\ \mathbf{X}^T)$. Suppose the log-conditional density function of y given $(\mathbf{X}^{\mathrm{T}},\ U)$ is

$$C_1(\boldsymbol{\phi}) f\left(m(\mathbf{X}^T, U)\right) + C_2(y,\ \boldsymbol{\phi}), \tag{3.1}$$

where $f(\cdot,\ \cdot)$ is a known function. $\boldsymbol{\phi}$ is a vector of unknown nuisance parameters independent of $m(\mathbf{X}^T, U)$, therefore the dispersion function $C_1(\boldsymbol{\phi}) > 0$ would not affect the maximum likelihood property of $m(\mathbf{X}^T, U)$. Denote by $\boldsymbol{\beta}(U) = (\beta_1(U), \cdots, \beta_p(U))^T$ the p-dimensional unknown functional vector holding the varying coefficient functions. The mean regression function is supposed to be linear via an unknown monotonic link function $g(\cdot)$ as

$$g\left\{m(\mathbf{X}^T, U)\right\} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U). \tag{3.2}$$

To make the model identifiable for the proposed Maximum Rank Correlation Estimation (MRCE) method, there is a constrain condition on the varying coefficients. Denote the norm of $\boldsymbol{\beta}(\cdot)$ at $U = 0$ by $||\boldsymbol{\beta}(0)||$. Assume $\|\boldsymbol{\beta}(0)\|$ is known in advance. The norm of $\boldsymbol{\beta}(\cdot)$ at $U = 0$ does not impact upon the monotonicity of the link function $g(\cdot)$, as $||\boldsymbol{\beta}(0)|| > 0$. However, it does impose identifiability issue for the MRCE method to be proposed. The reason is in that the proposed MRCE method calculates the integral of the first order derivative of $\|\boldsymbol{\beta}(\cdot)\|$. However, there is no information of $||\boldsymbol{\beta}(0)||$. For simplicity, assume $\|\boldsymbol{\beta}(0)\| = 1$. Thus, in this thesis, the MRCE method in fact estimates quantities that are proportional to the true varying coefficients with a positive factor $||\boldsymbol{\beta}(0)||$.

To make further presentation clear, some important notions are provided here. For any given $U = u$, let $\boldsymbol{\beta}_0(u) = \frac{\boldsymbol{\beta}(u)}{\|\boldsymbol{\beta}(u)\|}$ denotes the direction of the varying coefficient vector, and $\|\boldsymbol{\beta}(u)\| = \left\{\boldsymbol{\beta}(u)^{\mathrm{T}}\boldsymbol{\beta}(u)\right\}^{1/2}$ is the L_2 norm that represents the length of the varying coefficients. Throughout this thesis, for any function $f(\cdot)$, $\dot{f}(\cdot)$ is used to denote its derivative.

3.2 Application of Maximum Rank Correlation

Motivated by the potential consequences of model miss-specification, Han (1987) introduced the Maximum Rank Correlation (MRC) estimator for generalized transformation models for which the transformation function and error distribution are both unknown. The estimator is $n^{1/2}$ -consistent and asymptotically normal (Sherman, 1993). Estimation of the transformation function is further discussed by Horowitz (1996), Ye and Duan (1997), Cheng (2002) and Zhou *et al.* (2009). Lian and Peng (2013) utilized the idea of MRC estimation for linear transformation regression models. Inspired by these works, the thesis intends to extend the application of MRC for generalized varying coefficient models. Without loss of generality, this thesis only considers strictly increasing link functions.

Given $\mathbf{X}_i^T\boldsymbol{\beta}(U_i) \geq \mathbf{X}_j^T\boldsymbol{\beta}(U_j)$, the monotonicity of $g(\cdot)$ and the independence of U and $\mathbf{X}$ ensure that

$$P(y_i \geq y_j|\mathbf{X}_i^T, U_i, \mathbf{X}_j^T, U_j) \geq P(y_i \leq y_j|\mathbf{X}_i^T, U_i, \mathbf{X}_j^T, U_j). \tag{3.3}$$

Since terms for which $i = j$ make a negligible asymptotic contribution, and ties in (3.3) are irrelevant to our interests (Sherman, 1993), (3.3) can be relaxed to

$$P(y_i > y_j|\mathbf{X}_i^T, U_i, \mathbf{X}_j^T, U_j) > P(y_i < y_j|\mathbf{X}_i^T, U_i, \mathbf{X}_j^T, U_j), \tag{3.4}$$

when $\mathbf{X}_i^T\boldsymbol{\beta}(U_i) > \mathbf{X}_j^T\boldsymbol{\beta}(U_j)$ and $i \neq j$. From (3.4), the global rank correlation function can be constructed as

$$\sum_{i \neq j} I(y_i > y_j) I\Big(\mathbf{X}_i^T\boldsymbol{\beta}(U_i) > \mathbf{X}_j^T\boldsymbol{\beta}(U_j)\Big), \tag{3.5}$$

where $I(\cdot)$ stands for the indicator function

$$\begin{aligned} I(\cdot) &= 1, \quad \text{if } (\cdot) \text{ is ture;} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The primary intention of this thesis is to estimate the varying coefficients through application of the rank correlation function with local regression techniques. One has to be cautious when proceeding to local regression techniques. At any location $U = u$, with local linear approximation to the varying coefficients

$$\boldsymbol{\beta}(U_i) \approx \boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_i - u),$$

the localized version of (3.5) is given by

$$\begin{aligned} &\sum_{i \neq j} I(y_i > y_j) I\left(\mathbf{X}_i^T[\boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_i - u)] > \mathbf{X}_j^T[\boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_j - u)]\right) \\ &\times K_h(U_i - u) K_h(U_j - u), \end{aligned} \tag{3.6}$$

where $K_h(t) = \frac{K(t/h)}{h}$, $K(t)$ is the kernel function, and h is the smoothing parameter defining the bandwidth at $U = u$. Potential estimator $\hat{\boldsymbol{\beta}}(u)$ and $\hat{\dot{\boldsymbol{\beta}}}(u)$ are those that maximize the objective function (3.6). However, it is recognized that (3.6) is not identifiable.

Suppose $\hat{\boldsymbol{\beta}}(u)$ and $\hat{\dot{\boldsymbol{\beta}}}(u)$ constitute the maximizer of (3.6). For any positive number K, multiply K to this maximizer gives $\hat{\mathbf{a}}(u) = K\hat{\boldsymbol{\beta}}(u)$ and $\hat{\mathbf{a}}(u) = K\hat{\dot{\boldsymbol{\beta}}}(u)$. Substitute $\hat{\mathbf{a}}(u)$ and $\hat{\mathbf{a}}(u)$ into (3.6), the locally weighted rank correlation function does not change, since

$$
\begin{aligned}
& \sum_{i \neq j} I(y_i > y_j) I\left(\mathbf{X}_i^T[\boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_i - u)] > \mathbf{X}_j^T[\boldsymbol{\beta}(u) + \dot{\boldsymbol{\beta}}(u)(U_j - u)]\right) \\
& \times K_h(U_i - u) K_h(U_j - u) \\
= & \sum_{i \neq j} I(y_i > y_j) I\left(\mathbf{X}_i^T[\mathbf{a}(u) + \dot{\mathbf{a}}(u)(U_i - u)] > \mathbf{X}_j^T[\mathbf{a}(u) + \dot{\mathbf{a}}(u)(U_j - u)]\right) \\
& \times K_h(U_i - u) K_h(U_j - u).
\end{aligned}
$$

That is to say, $\hat{\mathbf{a}}(u)$ and $\hat{\dot{\mathbf{a}}}(u)$ also constitute a maximizer of the local rank correlation function. Therefore, the rank correlation function is not identifiable.

Direct application of the rank correlation idea is implausible. To conquer this issue, the thesis proposes to estimate the varying coefficient function $\boldsymbol{\beta}(\cdot)$ in two stages. The intention here is to solve the identifiability issue in rank correlation functions. Firstly, estimate the directions of the varying coefficients, $\hat{\boldsymbol{\beta}}_0(\cdot)$; secondly, estimate the norm of the varying coefficients, $||\hat{\boldsymbol{\beta}}(\cdot)||$. Estimators of the varying coefficients are therefore composed by $\hat{\boldsymbol{\beta}}_0(\cdot)$ and $||\hat{\boldsymbol{\beta}}(\cdot)||$. Given the estimated varying coefficients, $\hat{\boldsymbol{\beta}}(\cdot)$, the unknown link function can be accessed afterwards.

3.3 Estimation Procedure

Estimation of $\boldsymbol{\beta}(\cdot)$ and the unknown monotonic link function $g(\cdot)$ are the two main goals of this study. With the functional link function $g(\cdot)$ given, the problem of estimating the varying coefficients is straightforward. The Local Maximum Likelihood Estimation (Local MLE) method would provide good estimates. Since the link function is totally unknown, exploring the structure of the varying coefficients becomes complicated. By assuming that the link function is monotonic, the thesis looks for a possible path to access the varying coefficients.

3.3.1 One-Step Estimation of $\boldsymbol{\beta}_0(\cdot)$

Without loss of generality, it is assumed that the link function $g(\cdot)$ is strictly increasing. Thus, the proposed estimation procedure which is based on maximum rank correlation is possible. The rank correlation between y and $\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U)$ is defined in (3.5) as

$$\sum_{i\neq j} I(y_i > y_j) I\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}(U_i) > \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}(U_j)\right).$$

At any $U = u$, instead of local linear approximation to the varying coefficients, the local constant approximation gives $\boldsymbol{\beta}(U_i) \approx \boldsymbol{\beta}(u)$. The local rank correlation function is then defined as

$$\sum_{i\neq j} I(y_i > y_j) I\left(\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}(u) > \mathbf{X}_j^{\mathrm{T}}\boldsymbol{\beta}(u)\right) K_{h_1}(U_i - u) K_{h_1}(U_j - u),$$

where $K_{h_1}(t) = \frac{K(t/h_1)}{h_1}$, and where $K(t)$ is the kernel function, and h_1 is the smoothing parameter defining the width of the neighbouring at $U = u$.

Let $\mathbf{X}_i^T\boldsymbol{\beta}(u) = \mathbf{X}_i^T\boldsymbol{\beta}_0(u)\|\boldsymbol{\beta}(u)\|$, and denote by $\mathbf{a} = \boldsymbol{\beta}_0(u)$ the vector that holds directions of the varying coefficients at any given u. With $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$, the objective function for estimation of $\boldsymbol{\beta}_0(u)$ is constructed as

$$\begin{aligned} L(\mathbf{a}) &= \sum_{i\neq j} I(y_i > y_j) I\left(\mathbf{X}_i^T\boldsymbol{\beta}_0(u)\|\boldsymbol{\beta}(u)\| > \mathbf{X}_j^T\boldsymbol{\beta}_0(u)\|\boldsymbol{\beta}(u)\|\right) \\ &\quad \times K_{h_1}(U_i - u)K_{h_1}(U_j - u) \\ &= \sum_{i\neq j} I(y_i > y_j) I\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{a} > \mathbf{X}_j^{\mathrm{T}}\mathbf{a}\right) K_{h_1}(U_i - u)K_{h_1}(U_j - u). \quad (3.7) \end{aligned}$$

The problem now is how to find maximizers of the objective function (3.7). This thesis is to introduce two possible solutions. We call them Method 1 and Method 2 for simplicity.

3.3.2 Method 1: General estimation

The objective function $L(\mathbf{a})$ is not continuous due to the indicator function $I\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{a} > \mathbf{X}_j^{\mathrm{T}}\mathbf{a}\right)$. For the practical purpose of utilizing mathematical algorithms, like the Newton-Raphson maximization/minimization algorithm, the optimization of $L(\mathbf{a})$ needs an extensive smooth approximation. Inspired by Lin and Peng (2013), a smoothing distribution function $\Phi(\cdot)$ is used to approximate the indicator function. The smoothing distribution function is defined as

$$\Phi(t) = \int_{-\infty}^{t} \phi(u)du, \text{ and } \phi(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}.$$

Therefore,

$$I\left(\mathbf{X}_i^{\mathrm{T}}\mathbf{a} > \mathbf{X}_j^{\mathrm{T}}\mathbf{a}\right) \approx \Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right),$$

where δ is a tuning parameter that controls the functional pattern of $\Phi(\cdot)$. As the sample size increases, for any given positive constant $\delta \to 0_+$, the smoothing distribution function satisfies that

$$\Phi\left(\frac{(\mathbf{X}_i-\mathbf{X}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right)\to 1,\ \text{if}\ \mathbf{X}_i^{\mathrm{T}}\mathbf{a}>\mathbf{X}_j^{\mathrm{T}}\mathbf{a},$$

and

$$\Phi\left(\frac{(\mathbf{X}_i-\mathbf{X}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right)\to 0,\ \text{if}\ \mathbf{X}_i^{\mathrm{T}}\mathbf{a}<\mathbf{X}_j^{\mathrm{T}}\mathbf{a}.$$

This ensures that the smoothing approximation makes sense. The objective function (3.7) can therefore be approximated by

$$L(\mathbf{a})=\sum_{i\neq j}I(y_i>y_j)\Phi\left(\frac{(\mathbf{X}_i-\mathbf{X}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right)K_{h_1}(U_i-u)K_{h_1}(U_j-u),$$
$$\text{with}\quad \mathbf{a}^{\mathrm{T}}\mathbf{a}=1. \qquad (3.8)$$

In (3.8), there is no identifiability issue, as long as $\mathbf{a}$ is constrained by setting the norm of $\mathbf{a}$ to be equal to 1. If $\hat{\mathbf{a}}$ maximizes $L(\mathbf{a})$, $\hat{\mathbf{a}}$ is an estimator of $\boldsymbol{\beta}_0(u)$, and is denoted by $\hat{\boldsymbol{\beta}}_0(u)$.

There is one important issue to be noted here. For the application of local constant approximation to the varying coefficients, the intercept term is not added into the model. Suppose one wants to allow the linear combination to have an intercept term. Denote the covariates, the varying coefficients and their directions as $\mathbf{Z}=(1,\mathbf{X}^T)^T$, $\boldsymbol{\alpha}(U)=(\alpha(U),\boldsymbol{\beta}^T(U))^T$, and $\boldsymbol{\alpha}_0(U)=(\alpha_0(U),\boldsymbol{\beta}_0^T(U))^T$, respectively, where $\alpha(\cdot)$ and $\alpha_0(\cdot)$ are the intercept term and its direction, and $\|\boldsymbol{\alpha}(0)\|=1$. The link transformation gives

$$\begin{aligned} g\Big\{m(\mathbf{Z},U)\Big\} &= \mathbf{Z}^T\boldsymbol{\alpha}(U)\\ &= \alpha(U)+\mathbf{X}^T\boldsymbol{\beta}(U)\\ &= \big(\alpha_0(U)+\mathbf{X}^T\boldsymbol{\beta}_0(U)\big)\,||\boldsymbol{\alpha}(U)||. \end{aligned}$$

At any $U = u$, the objective function (3.8) is given by

$$\begin{aligned} & \sum_{i \neq j} I(y_i > y_j) \Phi \left(\frac{(\mathbf{Z}_i - \mathbf{Z}_j)^{\mathrm{T}} \boldsymbol{\alpha}(u)}{\delta} \right) K_{h_1}(U_i - u) K_{h_1}(U_j - u) \\ = \ & \sum_{i \neq j} I(y_i > y_j) \Phi \left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \boldsymbol{\beta}_0(u)}{\delta} \right) K_{h_1}(U_i - u) K_{h_1}(U_j - u). \end{aligned}$$

It is clear that the direction of the intercept term is cancelled out in the maximization, which makes the estimation impossible. Indeed, one may wish to find an approach that allows the existence of an intercept term. This thesis suggests to consider this question in further research.

3.3.3 Method 2: Estimation of directions with strictly positive (negative) component

In certain occasions, some components of the covariates impact on the response variable positively or negatively. For example, some significant factors that lead to specific diseases are always positively related to the number of patients suffering from such diseases, while accessibility of medication on the contrary is negatively associated with the number of patients. In such cases, where at least one of the components of the varying coefficients is strictly positive or negative, one can get rid of the identifiability issue in the stage of direction estimation. For simplicity, this thesis only considers the case that at least one of the components of the varying coefficient vector is strictly positive.

Without loss of generality, suppose the first component of the varying coefficients is strictly positive, *i.e.*

$$\beta_1(\cdot) > 0.$$

At any given $U = u$, let $\mathbf{a} = \boldsymbol{\beta}_0(u)$ and $a_1 > 0$ be the first component of vector $\mathbf{a}$. Then

$$\mathbf{a} = a_1 * \mathbf{a}^\star,$$

given that

$$\mathbf{a}^\star = \left(1, \frac{a_2}{a_1}, \cdots, \frac{a_p}{a_1}\right)^T$$

is achieved by dividing the remaining $p-1$ components of $\mathbf{a}$ by the first component a_1. The objective function (3.7) can be transformed into

$$L(\mathbf{a}^\star) = \sum_{i \neq j} I(y_i > y_j) I\left(\mathbf{X}_i^{\mathrm{T}} \mathbf{a}^\star > b X_j^{\mathrm{T}} \mathbf{a}^\star\right) K_{h_1}(U_i - u) K_{h_1}(U_j - u).$$

Note that, as the first component of $\mathbf{a}^\star$ is 1, the identifiability issue in the maximization of rank correlation is eliminated. In the same vein as Method 1 which aims at a more general estimation, replace the indicator function by a smoothing distribution. Denote

$$I\left(\mathbf{X}_i^{\mathrm{T}} \mathbf{a}^\star > \mathbf{X}_j^{\mathrm{T}} \mathbf{a}^\star\right) \approx \Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \mathbf{a}^\star}{\delta}\right),$$

where

$$\Phi(t) = \int_{-\infty}^{t} \phi(u) du, \text{ and } \phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2},$$

and δ is a tuning parameter that controls the functional pattern of $\Phi(\cdot)$. As the sample size increases, for any given positive constant $\delta \to 0_+$,

$$\Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \mathbf{a}^\star}{\delta}\right) \to 1, \quad \text{if } \mathbf{X}_i^{\mathrm{T}} \mathbf{a}^\star > \mathbf{X}_j^{\mathrm{T}} \mathbf{a}^\star,$$

and

$$\Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \mathbf{a}^\star}{\delta}\right) \to 0, \quad \text{if } \mathbf{X}_i^{\mathrm{T}} \mathbf{a}^\star < \mathbf{X}_j^{\mathrm{T}} \mathbf{a}^\star.$$

The objective local rank correlation function is then constructed as

$$L(\mathbf{a}^\star) = \sum_{i \neq j} I(y_i > y_j) \Phi \left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \mathbf{a}^\star}{\delta} \right) K_{h_1}(U_i - u) K_{h_1}(U_j - u). \quad (3.9)$$

Suppose $\hat{\mathbf{a}}^\star$ is the maximizer of (3.9). Then the estimator $\hat{\boldsymbol{\beta}}_0(\cdot)$ is obtained by standardizing $\hat{\mathbf{a}}^\star$, *i.e.*

$$\hat{\boldsymbol{\beta}}_0(\cdot) = \frac{\hat{\mathbf{a}}^\star}{\|\hat{\mathbf{a}}^\star\|}.$$

Comparing with Method 1, in Method 2, the dimension of unknown directions is reduced from p to $p-1$. Therefore, there are relatively more information locally in the maximization iteration procedure. It is potential that Method 2 would give better estimators for the directions than that of Method 1.

3.3.4 Two-Step Estimation of $\boldsymbol{\beta}_0(\cdot)$

In the maximum rank correlation approach, the varying coefficients are estimated by a multiplication of the estimated varying coefficients in the unit circle and the estimates of the norm. When the directions of varying coefficients $\boldsymbol{\beta}_j(\cdot)$, $j = 1, \cdots p$, have relatively similar degrees of smoothness, the one-step approach proposed suffices. We may get reasonable estimates for directions of the varying coefficients by applying Method 1 generally, or by adopting Method 2 when the varying-coefficient vector includes a strictly positive component. However, the estimated curve of the directions could be either over or under smoothed, if the differences in smoothness amongst the directions of the varying coefficients are not negligible. In such cases, a more sophisticated way is to estimate the directions of the varying coefficients with a two-step estimation procedure.

The idea of two-step estimation is clearly explained in the literature of Fan and Zhang (1999). However, the situation in this context is different, as a two-step estimation only occurs at the stage of searching for directions of the varying coefficients. The intuition is to use a smaller bandwidth in the first stage using the one-step estimation method proposed. This would provide an initial estimator for the directions which have smaller bias and larger variance. As local linear smoothing would not impact on the bias site, but improves the performance on the variance site. In the second stage, treat the smoother initial estimator with a local linear modelling to reduce the variance.

Without loss of generality, in model (3.1), assume that the direction of the first varying coefficient function $\beta_1(\cdot)$ is smoother than that of other varying coefficient functions. Consider the objective function (3.8), suppose now that estimates of the directions with one-step estimation method and bandwidth h_1 are derived. The estimators are denoted as $\hat{\beta}_{0j}(\cdot)$, $j = 2, \cdots, p$. The thesis now introduces the two-step estimation method for $\beta_{01}(\cdot)$. The two-step estimation method involves two stages:

- In the first stage of the two-step estimation, apply the one-step estimation method with bandwidth h_{20}. h_{20} is smaller comparing to the one-step method bandwidth h_1. This smaller bandwidth would yield initial estimates $\tilde{\beta}_{01}(U_i)$ of $\beta_{01}(U_i)$, $i = 1, \cdots, n$. The initial estimator has smaller bias and larger variance than that of a standard one-step estimation using a larger bandwidth h_1.

- In the second stage of the two-step estimation, use a local linear approach to correct the variance site. Consider $\tilde{\beta}_{01}(U_i)$, $i = 1, \cdots, n$, as a sample observation drawn from a linear model

$$\eta = \beta_{01}(U) + \epsilon. \tag{3.10}$$

At any given point $U = u$, a standard local linear modelling supplies the two-step estimator

$$\hat{\beta}_{01}(u) = (1,\ 0)\left(\mathbf{V}^{\mathrm{T}} W \mathbf{V}\right)^{-1} \mathbf{V}^{\mathrm{T}} W \boldsymbol{\eta},$$

where

$$\mathbf{V} = \begin{pmatrix} 1 & U_1 - u \\ \vdots & \vdots \\ 1 & U_n - u \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \tilde{\beta}_{01}(U_1) \\ \vdots \\ \tilde{\beta}_{01}(U_n) \end{pmatrix},$$

and

$$W = \mathrm{diag}\left(K_{h_2}(U_1 - u),\ \cdots,\ K_{h_2}(U_n - u)\right).$$

h_2 is the bandwidth used in the second stage of the estimation. $K_{h_2}(t) = \frac{K(t/h_2)}{h_2}$ is the weight function, where $K(t)$ is the kernel function.

Up to this line, the two-step estimation method is only applied to the first component of the direction of the varying coefficient vector. It is intuitive that, the two-step estimation would improve on the estimation of the direction of the first component, while, the directions of the remaining $p-1$ varying coefficients are estimated with one-step method using bandwidth h_1. Since the directions of the varying coefficients are estimated using different methods, the norm of $\hat{\beta}_0(\cdot)$ is no longer equal to one. Therefore, a two-step estimation method should be ended by a standardization of the estimated directions provided by one-step method and two-step method.

Naturally, the next question following the introduction of two-step estimation method would be how to identify the smoothness of the functional directions of the varying coefficients. This question is interesting and complicated in the real world. Luckily, it may not be needed to answer this question in this thesis. Since a two-step estimator would not be worse than a one-step estimator (Fan and Zhang, 1996), two-step estimation method can be extended to those rougher functional directions of the varying coefficients as well. One should bear in mind that when two-step estimation method is applied for all components of the varying coefficient vector, there are more second stage bandwidths included into the estimation. There will be a sacrifice in computation cost, especially when bandwidths are derived by data-driven algorithms.

3.3.5 Estimation of $\|\boldsymbol{\beta}(\cdot)\|$

Denote the norm of the varying coefficient vector $\|\boldsymbol{\beta}(\cdot)\|$ by $N(\cdot)$. Let

$$z = \mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0(U), \quad \text{and } z_i = \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0(U_i),$$

where $\hat{\boldsymbol{\beta}}_0(U)$ is the estimator of the directions of the varying coefficients by either one-step or two-step estimation method.

Replacing the directions of the varying coefficients by their estimators gives

$$\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0(U_i)N(U_i) \approx \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_0(U_i)N(U_i),$$

which yields the following rank correlation between y and $zN(U)$:

$$\sum_{i\neq j} I(y_i > y_j) I\left(z_i N(U_i) > z_j N(U_j)\right).$$

For any given u, given U_i is in a small neighbourhood of u, the Taylor's expansion leads to

$$N(U_i) \approx N(u) + \dot{N}(u)(U_i - u).$$

The local rank correlation is then approximated by

$$\sum_{i \neq j} I(y_i > y_j) I\left(z_i \left\{N(u) + \dot{N}(u)(U_i - u)\right\} > z_j \left\{N(u) + \dot{N}(u)(U_j - u)\right\}\right) \\ \times K_{h_n}(U_i - u) K_{h_n}(U_j - u),$$

where $K_{h_n}(t) = \frac{K(t/h_n)}{h_n}$, with $K(t)$ is the kernel function, and h_n is the smoothing parameter defining the width of the neighbouring at $U = u$. Because $N(u) > 0$, the above objective function is equivalent to

$$\sum_{i \neq j} I(y_i > y_j) I\left(z_i \left\{1 + c(u)(U_i - u)\right\} > z_j \left\{1 + c(u)(U_j - u)\right\}\right) \\ \times K_{h_n}(U_i - u) K_{h_n}(U_j - u), \quad (3.11)$$

where $c(u)$ corresponds to $\dot{N}(u)/N(u)$. If $\hat{c}(u)$ maximise (3.11), $\hat{c}(u)$ is an estimator of $\dot{N}(u)/N(u)$, and the estimator of $N(u)$ is generated by

$$\hat{N}(u) = \exp\left\{\int_0^u \hat{c}(u) du\right\}. \quad (3.12)$$

The estimator of $\boldsymbol{\beta}(u)$ is therefore conducted via

$$\hat{\boldsymbol{\beta}}(u) = \hat{N}(u)\hat{\boldsymbol{\beta}}_0(u).$$

One may wish to search for $\hat{c}(u)$ by maximizing a smoothing approximation of (3.11). Similar to that in estimating the directions of the varying coefficients, the objective function can be defined as

$$\sum_{i\neq j} I(y_i > y_j)\Phi\left(\frac{z_i\{1+c(u)(U_i-u)\}-z_j\{1+c(u)(U_j-u)\}}{\delta}\right) \times K_{h_n}(U_i-u)K_{h_n}(U_j-u). \quad (3.13)$$

A standard Newton-Raphson maximization provides the maximizer $\hat{c}(u)$, and hence the estimator $\hat{N}(u)$. However, this intuition confronts technical difficulties. Simulations indicates that Newton-Raphson maximization/minimization method can hardly provide proper estimates for $c(\cdot)$. Therefore, we have to abandon this idea and enquire grid regression method to find the maximizer of (3.11).

3.3.6 Estimation of $g(\cdot)$

Estimation of the link function $g(\cdot)$ is in fact about the estimation of the conditional mean regression function $m(\mathbf{X}, U)$. The link transformation $g\left\{m(\mathbf{X}^T, U)\right\} = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U)$ is equivalent to $m(\mathbf{X}^T, U) = g^{-1}\left\{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U)\right\}$. Without loss of generality, use $g(\cdot)$ to denote the inverse of the link function, *i.e.* $m(\mathbf{X}^T, U) = g\left\{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U)\right\}$, which pertains the quality of monotonicity. Once the estimator of $\boldsymbol{\beta}(\cdot)$ is obtained, the estimation of $g(\cdot)$ becomes easier. Firstly, apply the local maximum likelihood estimation to get an initial estimator $\tilde{g}(\cdot)$ of $g(\cdot)$; secondly, make use of the monotonicity of $g(\cdot)$ to refine the initial estimator to get the final estimator $\hat{g}(\cdot)$. The details are as follows.

Let $t_i = \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}(U_i)$. For any given t, denote the linear approximation of $g(t_i)$ by

$$g(t) + \dot{g}(t)(t_i - t).$$

By simple calculation, following local log-likelihood function of $g(t)$ and $\dot{g}(t)$

can be derived, *i.e.*

$$C_1(\boldsymbol{\phi})\sum_{i=1}^{n} f\left(d + q(t_i - t),\ y_i\right) K_{h_l}(t_i - t) + \sum_{i=1}^{n} C_2(y_i, \boldsymbol{\phi}) K_{h_l}(t_i - t). \quad (3.14)$$

where d and q represents $g(t)$ and $\dot{g}(t)$ respectively, and h_l is the bandwidth used. An initial estimator $\tilde{g}(t)$ of $g(t)$ is the part, corresponding to d of maximiser of (3.14).

Denote the longest subset of $\{(t_i,\ \tilde{g}(t_i)) :\ i = 1,\ \cdots,\ n\}$ that satisfies the monotonicity,

$$t_{(i)} \leq \cdots \leq t_{(T)} \quad \text{and} \quad \tilde{g}(t_{(i)}) \leq \cdots \leq \tilde{g}(t_{(T)}),$$

as $\left\{\left(t_{(i)},\ \tilde{g}(t_{(i)})\right) :\ i = 1,\ \cdots,\ T\right\}$, and treat this subset as a sample from the following univariate non-parametric regression model

$$\eta = g(\xi) + \epsilon. \quad (3.15)$$

By applying the standard local linear modelling, for any given t, the estimator $\hat{g}(t)$ of $g(t)$ is given by

$$\hat{g}(t) = (1,\ 0)\left(\mathbf{T}^{\mathrm{T}} W \mathbf{T}\right)^{-1} \mathbf{T}^{\mathrm{T}} W \boldsymbol{\eta},$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & t_{i_1} - t \\ \vdots & \vdots \\ 1 & t_{i_T} - t \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \tilde{g}(t_{i_1}) \\ \vdots \\ \tilde{g}(t_{i_T}) \end{pmatrix},$$

and

$$W = \operatorname{diag}\left(K_{h_l}(t_{i_1} - t),\ \cdots,\ K_{h_l}(t_{i_T} - t)\right).$$

When initial estimator $\tilde{g}(t)$ is achieved, if the length of the longest subset

$$\left\{\left(t_{(i)},\ \tilde{g}(t_{(i)})\right):\ i=1,\ \cdots,\ T\right\}$$

is close to the sample size n, one may simply achieve the estimator of the link function by

$$\hat{g}(t) = O(\tilde{g}(t)|t)$$

where $O(\tilde{g}(t)|t)$ is a strictly increasing function obtained by sorting $\tilde{g}(t)$ along t.

3.4 Computational Algorithm

The employability of the proposed maximum rank correlation method depends on how the two hurdles: the identifiability issue and the computational maximization of rank correlation, are crossed. The former is conquered via estimation of the directions $\boldsymbol{\beta}_0(\cdot)$ and the norm $N(\cdot)$ of the varying coefficients separately, where the estimator of the varying coefficient vector is composed through $\boldsymbol{\beta}(\cdot) = \boldsymbol{\beta}_0(\cdot)N(\cdot)$. This thesis votes for Newton-Raphson algorithm, which is an ideal tool that provides proper estimators. For the latter obstacle, the author attempts to replace the indicator function by its smooth approximation, so that Newton-Raphson maximization algorithm (see Section 5.2 in Conte and De Boor, 1980) is applicable.

3.4.1 One Step Estimation of $\boldsymbol{\beta}_0(U)$

This thesis proposes two possible ways of obtaining estimators of $\boldsymbol{\beta}_0(U)$, which are briefly named Method 1 and Method 2.

Method 1

Recall the objective function (3.8). A penalty function is added to (3.8) and leads to the following practical objective function

$$\ell(\mathbf{a}) = \sum_{i \neq j} I(y_i > y_j)\Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right) K_{h_1}(U_i - u)K_{h_1}(U_j - u) - \lambda\mathbf{a}^{\mathrm{T}}\mathbf{a}, \tag{3.16}$$

where δ and λ are tuning parameters, and $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$. At each $U = u$, maximization of (3.16) with respect to $\mathbf{a}$ provides the estimator $\hat{\boldsymbol{\beta}}_0(u)$ of $\boldsymbol{\beta}_0(u)$. The motivation of applying the penalty function here is that it helps to deal with the convergence problem during the iterative Newton-Raphson maximization. Without the control of such penalty function, it is not uncommon that the iteration of updating estimates is hard to converge.

Let $\mathbf{a}_n$ be the current $\mathbf{a}$. Denote the first two orders of derivatives of the practical objective function by

$$\begin{aligned}
\dot{\ell}(\mathbf{a}_n) &= \frac{1}{\delta}\sum_{i\neq j} I(y_i > y_j)\phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}_n}{\delta}\right)(\mathbf{X}_i - \mathbf{X}_j) \times \\
&\quad K_{h_1}(U_i - u)K_{h_1}(U_j - u) - 2\lambda\mathbf{a}_n \\
&= \frac{1}{\sqrt{2\pi}\delta}\sum_{i\neq j} I(y_i > y_j)\exp\left(-\frac{\left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}_n\right\}^2}{2\delta^2}\right)(\mathbf{X}_i - \mathbf{X}_j) \times \\
&\quad K_{h_1}(U_i - u)K_{h_1}(U_j - u) - 2\lambda\mathbf{a}_n,
\end{aligned}$$

and

$$\begin{aligned}
\ddot{\ell}(\mathbf{a}_n) &= \frac{1}{\delta^2}\sum_{i\neq j} I(y_i > y_j)\dot{\phi}\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}_n}{\delta}\right)K_{h_1}(U_i - u)K_{h_1}(U_j - u) \times \\
&\quad (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} - 2\lambda I_p \\
&= \frac{1}{\sqrt{2\pi}\delta^3}\sum_{i\neq j} I(y_i > y_j)\exp\left(-\frac{\left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}_n\right\}^2}{2\delta^2}\right)(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}_n \times \\
&\quad K_{h_1}(U_i - u)K_{h_1}(U_j - u)(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} - 2\lambda I_p,
\end{aligned}$$

where I_p is an identity matrix of size p.

There are two possible algorithms for updating the maximizer. (1) Update $\mathbf{a}$ through

$$\mathbf{a}_{n+1} = \mathbf{a}_n - \frac{\ddot{\ell}(\mathbf{a}_n)}{\dot{\ell}(\mathbf{a}_n)} \tag{3.17}$$

until convergence. The condition $\mathbf{a}^T\mathbf{a} = 1$ is not considered in the progress. When (3.17) converges to the maximizer $\hat{\mathbf{a}}$, standardization of $\hat{\mathbf{a}}$ gives the estimator of $\hat{\boldsymbol{\beta}}_0(\cdot)$.

(2) Conduct the updating process as

$$\mathbf{a}_{n+1} = \frac{\mathbf{a}_n - \ddot{\ell}(\mathbf{a}_n)^{-1}\dot{\ell}(\mathbf{a}_n)}{\|\mathbf{a}_n - \ddot{\ell}(\mathbf{a}_n)^{-1}\dot{\ell}(\mathbf{a}_n)\|}. \tag{3.18}$$

The condition $\mathbf{a}^T\mathbf{a} = 1$ is considered in each and every step of the updating iteration. Both algorithms give proper estimates, while the first one is computationally cheaper.

To provide an initial value $\mathbf{a}_0$ for the Newton-Raphson algorithm, simply pretend that $(\mathbf{X}_i^{\mathrm{T}},\ U_i,\ y_i)$, $i = 1,\ \cdots,\ n$, are natural observations from the varying-coefficient model

$$y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}(U) + \epsilon,$$

and proceed to a local linear approach. At any $U = u$, the estimator $\tilde{\boldsymbol{\beta}}(\cdot)$ is constructed as

$$\tilde{\beta}_i(u) = {e_{2i+1}}^T \left(X_0^{\mathrm{T}} W_0 X_0\right)^{-1} X_0^{\mathrm{T}} W_0 Y, \quad i \in (1, \dots, p),$$

where

$$X_0 = \begin{pmatrix} X_{11} & X_{11}(U_1-u) & \dots & X_{p1} & X_{p1}(U_1-u) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1n} & X_{1n}(U_n-u) & \dots & X_{pn} & X_{pn}(U_1-u) \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

and

$$W_0 = \operatorname{diag}\left(K_{h_0}(U_1-u),\ \cdots,\ K_{h_0}(U_n-u)\right),$$

and h_0 is the initial bandwidth that controls the span of the locality. Once the estimates $\tilde{\boldsymbol{\beta}}(u)$ is in hand, it is normalized by dividing by its norm $\|\tilde{\boldsymbol{\beta}}(u)\|$. Denote by $\tilde{\boldsymbol{\beta}}_0(u) = \tilde{\boldsymbol{\beta}}(u)/\|\tilde{\boldsymbol{\beta}}(u)\|$, and $\tilde{\boldsymbol{\beta}}_0(U)$ is used as the initial estimator for estimating $\boldsymbol{\beta}_0(u)$.

Method 2

Method 2 is built up under the assumption that the first component of the varying coefficients is strictly positive. Violation of this pre-assumption would invalidate the whole estimation procedure. Recall the objective function (3.9). Since it does not involve any constrains, direct application of Newton-Raphson maximization algorithm gives the maximizer of $\mathbf{a}^\star$.

Denote $\mathbf{a}^\star_{-1} = (a_2/a_1, \cdots, a_p/a_1)^T$. Let $\mathbf{a}^\star_{-1n}$ be the current $\mathbf{a}^\star_{-1}$ and denote the first two orders of derivatives of the practical objective function by

$$\begin{aligned}
\dot{\ell}(\mathbf{a}^\star_{-1n}) &= \frac{1}{\delta}\frac{1}{n(n-1)}\sum_{i\neq j} I(y_i > y_j)\phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}^\star_{-1n}}{\delta}\right) \\
&\quad \times K_{h_1}(U_i - u)K_{h_1}(U_j - u)(\mathbf{X}_i - \mathbf{X}_j) \\
&= \frac{1}{\sqrt{2\pi}\delta}\frac{1}{n(n-1)}\sum_{i\neq j} I(y_i > y_j)\exp\left(-\frac{\left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}^\star_{-1n}\right\}^2}{2\delta^2}\right) \\
&\quad \times K_{h_1}(U_i - u)K_{h_1}(U_j - u)(\mathbf{X}_i - \mathbf{X}_j),
\end{aligned}$$

and

$$\begin{aligned}
\ddot{\ell}(\mathbf{a}^\star_{-1n}) &= \frac{1}{\delta^2}\frac{1}{n(n-1)}\sum_{i\neq j} I(y_i > y_j)\dot{\phi}\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}^\star_{-1n}}{\delta}\right) \\
&\quad \times K_{h_1}(U_i - u)K_{h_1}(U_j - u)(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \\
&= \frac{1}{\sqrt{2\pi}\delta^3}\frac{1}{n(n-1)}\sum_{i\neq j} I(y_i > y_j)\exp\left(-\frac{\left\{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}^\star_{-1n}\right\}^2}{2\delta^2}\right) \\
&\quad \times (\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}\mathbf{a}^\star_{-1n}K_{h_1}(U_i - u)K_{h_1}(U_j - u)(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}}.
\end{aligned}$$

The updating algorithm of $\mathbf{a}^{\star}_{-1_{n+1}}$ is defined as

$$\mathbf{a}^{\star}_{-1_{n+1}} = \mathbf{a}^{\star}_{-1_{n}} - \ddot{\ell}(\mathbf{a}^{\star}_{-1_{n}})^{-1}\dot{\ell}(\mathbf{a}^{\star}_{-1_{n}}). \tag{3.19}$$

Update (3.19) until convergence would give the estimator $\hat{\mathbf{a}}^{\star}_{-1}$ of $\mathbf{a}^{\star}_{-1}$, and standardization of $\hat{\mathbf{a}}^{\star} = \left(1, \hat{\mathbf{a}}^{\star}_{-1}\right)^{T}$ leads to the estimator of $\hat{\boldsymbol{\beta}}_0(\cdot)$.

Simulation studies suggest that Method 2 depends heavily on the initial estimator. That is to say, initial estimates returned by treating a varying coefficient model is not sufficient. What this thesis suggests, is to let Method 1 provide initial estimates for Method 2. However, it is not necessary to proceed a full computation for Method 1. Update the estimates of Method 1 once or twice would give proper initial estimates for Method 2 without increasing too much on the computation site.

3.4.2 Two Step Estimation of $\boldsymbol{\beta}_0(\cdot)$

The initial stage of the two-step estimation shares the same computational algorithm as that of the one-step approach. The difference is in that this initial stage uses a smaller bandwidth. If h_1 is the optimal bandwidth for the one-step estimation, in the first stage of the two-step estimation, $h_{20} < h_1$ is used. Fan and Zhang (1999) suggest that a two-step estimator is not very sensitive to this initial bandwidth h_{20} as long as it ensures negligible bias. In practice, it is safe to multiply h_1 by a factor between 0 and 1 to get h_{20}. In this project, $h_{20} = 0.5h_1$ is applied for simplicity. On one hand, this first stage bandwidth h_{20} is not ridiculously small, thus proper first stage estimates of the directions of the varying coefficients are achievable. On the other hand, with this smaller bandwidth, the computational cost is not significantly increased.

Without loss of generality, suppose direction of the first component of the varying coefficients, $\beta_{01}(\cdot)$, is smoother than the rest of the unit circle varying coefficients. With the initial bandwidth h_{20}, one maximizes the target function (3.20)

$$\ell(\mathbf{a}) = \sum_{i \neq j} I(y_i > y_j) \Phi\left(\frac{(\mathbf{X}_i - \mathbf{X}_j)^{\mathrm{T}} \mathbf{a}}{\delta}\right) K_{h_{20}}(U_i - u) K_{h_{20}}(U_j - u), \quad (3.20)$$

with constrain condition $\mathbf{a}^{\mathrm{T}}\mathbf{a} = 1$. Either Method 1 or Method 2 (when applicable) may be applied. This process provides the initial estimator $\tilde{\beta}_{01}(\cdot)$ of $\beta_{01}(\cdot)$.

In the second stage of the two-step estimation, treat $\tilde{\beta}_{01}(U_i)$, $i = 1, \cdots, n$, as a realization from a linear model

$$\eta = \beta_{01}(U) + \epsilon,$$

where $\epsilon \sim N(0, \delta)$.

At any point $U = u$, a local linear approach produces the estimator $\bar{\beta}_{01}(\cdot)$ of $\beta_{01}(\cdot)$ as

$$\bar{\beta}_{01}(u) = e_1{}^T \left(X_2^{\mathrm{T}} W_2 X_2\right)^{-1} X_2^{\mathrm{T}} W_2 \boldsymbol{\eta},$$

where

$$X_2 = \begin{pmatrix} 1 & (U_1 - u) \\ \vdots & \vdots \\ 1 & (U_n - u) \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \tilde{\beta}_{01}(U_1) \\ \vdots \\ \tilde{\beta}_{01}(U_n) \end{pmatrix},$$

and

$$W_0 = \mathrm{diag}\left(K_{h_2}(U_1 - u), \ \cdots, \ K_{h_2}(U_n - u)\right).$$

Suppose the estimators $\bar{\beta}_{0j}(\cdot)$ $j = 2, \cdots p$ are obtained via the one-step estimation, and $\bar{\beta}_{01}(\cdot)$ is provided by the two-step estimation method, and denote $\|\bar{\boldsymbol{\beta}}_0(\cdot)\|$ as the norm of the estimator vector. The final estimators of $\beta_{0j}(\cdot)$ $j = 1, \cdots p$ are obtained by a standardization operation via

$$\hat{\beta}_{0j}(\cdot) = \frac{\bar{\beta}_{0j}(\cdot)}{\|\bar{\boldsymbol{\beta}}_0(\cdot)\|}.$$

In fact, the two-step estimation is not restricted to smoother functions. As Zhang and Fan (1999) have suggested, a two-step estimation mostly leads to increment in terms of approximation accuracy. This has been strongly supported by numerous simulation studies in later sections. Therefore, this thesis suggests to use two-step estimation method for each component of the varying coefficients for obtaining estimates of their directions. On one hand, this would improve the approximation accuracy for the directions of the varying coefficients. On the other hand, the time for identification of smoothness is saved.

In the work of Lin and Peng (2013), the norm of the varying coefficients is constrained to be $\|\boldsymbol{\beta}(\cdot)\| \equiv 1$. This naturally extends to a more general assumption, that is $\|\boldsymbol{\beta}(\cdot)\| = N(\cdot)$, where $N(\cdot)$ is some known positive function. When the norm is known in advance, the estimation of the varying coefficients reduces to the estimation of the unit circle varying coefficients. Consider the objective rank correlation function (3.21)

$$\sum_{i \neq j} I(y_i > y_j) I(\mathbf{X}_i N(U_i)\boldsymbol{\beta}_0(U_i) > \mathbf{X}_j N(U_j)\boldsymbol{\beta}_0(U_j)). \tag{3.21}$$

Local constant approximation to the unit circle varying coefficients at any u leads to

$$\sum_{i \neq j} I(y_i > y_j) I(\mathbf{X}_i N(U_i)\boldsymbol{\beta}_0(u) > \mathbf{X}_j N(U_j)\boldsymbol{\beta}_0(u)), \tag{3.22}$$

with ${\boldsymbol{\beta}_0}^T(u)\boldsymbol{\beta}_0(u) = 1$. Either the one-step or two-step method provides estimates of $\hat{\boldsymbol{\beta}}_0(U)$.

3.4.3 Estimation of the norm $\|\beta(\cdot)\|$

When the estimates of the directions of the varying coefficients are obtained, the estimation of the unknown norm becomes a univariate problem. The idea of grid regression is considered in the estimation of the unknown norm. Consider the following objective function defined in (3.11). This thesis tries to approach the maximizer with two algorithms, which is named the blind searching and the efficient searching in this context.

• **Blind searching**

At any point $U = u$, assume that the true value of $c(u)$ is within an interval $[C_0, C_N]$, where C_0 and C_N are two real numbers, and N is a positive integer determines the number of grid points to be searched. Divide this interval into N equally spaced subintervals. The dividing points are denoted as $C_0, C_1, \cdots, C_N$. A blind searching calculates the scores of the local rank correlation function for each of these dividing points. The estimator of $c(u)$ is chosen to be $\hat{c}(u) = C_k$, $k \in \{M, \cdots, N-M\}$ which maximizes the locally weighted rank correlation function, where $0 < M < N$ is a positive integer which ensures that the maximiser is not located at the edge of the searching interval, so that the region of interest, $[C_0, C_N]$, is properly defined. Once values of the estimator $\hat{c}(\cdot)$ at $U = U_i$, $i = 1, \cdots, n$ are fully achieved, the estimator for the unknown norm is given by

$$\hat{N}(u) = \exp\left\{\int_0^u \hat{c}(u)du\right\},$$

and the estimator of the varying coefficient vector is obtained via

$$\hat{\boldsymbol{\beta}}(u) = \hat{N}(u)\hat{\boldsymbol{\beta}}_0(u).$$

The idea of the grid regression is straightforward. However, the question comes to how large the searching region $[C_0, C_{N_c}]$ should be defined as a proper interval, and how much computational cost there would be if this interval is very ‘wide’. In the real world, there is no idea about how the true norm changes depending on U. Thus, the intuition is to proceed the searching in some ‘wide’ interval all the time for each $U = U_i$, $i = 1, \cdots, n$. When the searching interval is ‘wide’, it is unavoidable that the computation would be expensive. For instance, at point u, suppose the maximizer is at some point between -100 to 100. Calculation of the local rank correlation scores from -100 to 100 at grid points with subinterval length 0.1 results in 2001 computations, which is extremely expensive when sample size is large. Ironically, it is almost impossible to say a guessing searching interval is ‘wide’ enough. Practically, in terms of blind searching, reduction of the computational cost can only be made by assuming that a specified small searching interval to the best knowledge of researchers has already covered the location of the true maximizer.

Another issue to be highlighted is that some maximisers may hardly be accepted to be the estimators of $c(\cdot)$. At any point $U = u$, start the grid search from C_0 and moves toward C_N. If the rank correlation score firstly increases to some ‘peak’ point and then starts to decrease, and the score continues to decrease in a sufficient large number of steps, it is likely that the maximizer has already been identified and the maximization of the rank correlation occurs at the ‘peak’ point. Simulation studies show that around the 'peak' point, the score of the local rank correlation is not stable. There might be confusing cases which give wrong estimators of $c(\dot{})$. There might be a searching grid, at which the local rank correlation score is the highest. However, if one plots the local rank correlation scores against searching grids, one finds that this highest score is a sudden jump from the smooth curve of scores. These maximizers have to be avoid being treated as estimators of $c(\cdot)$ falsely.

• **Fast searching**

In simulation studies, the searching interval can be constructed around the true values of $c(\cdot)$. This would become a different story in the case of real data analysis. One has no information about where the true values of $c(\cdot)$ are. To let the maximization make sense, the scale of the grid searching region would not be set as some interval with very small width. That is to say, one has to set up a relatively wide searching interval, and the computation time could be frustrating. Therefore, a time-saving procedure for estimating $c(\cdot)$ which is free from the true values of $c(\cdot)$ and computationally cheap at the same time is desirable.

Simulation studies have confirmed that the rank correlation function (3.11) is quadratic. This is crucial to the idea of fast searching maximization to propose. Say, of interest, $c(u)$ is within an interval $[L_0, R_0]$, where L_0 and R_0 are two real numbers that can be very large. At any $U = u$m the maximiser of $c(u)$ is searched through the following procedures:

.1. First of all, instead of directly dividing the interval $[L_0, R_0]$ into numerous subintervals (for instance 1000 or even 10000), divide it into M equally spaced subintervals. In simulation studies, $M = 8$, $M = 10$ and $M = 12$ are used and all work well. Denote the grid points by $C1_0, C1_1, \cdots, C1_M$. Calculate the locally weighted rank correlation function at these $M+1$ nodes.

Suppose $C1_k$, $k \in \{1, \cdots, M-1\}$, is the value that maximizes the correlation function. Note that if $k = 0$ or $k = M$, it suggests that the initial definition of the searching interval $[L_0, R_0]$ is not properly defined and needs to be enlarged.

.2. Update the searching interval as $[L_1, R_1]$, where $L_1 = C1_{k-1}$ and $R_1 = C1_{k+1}$. Divide the updated searching interval into M equally spaced subintervals, and denote the grids as $C2_0, C2_1, \cdots, C2_M$. Calculate the locally weighted rank correlation function at these $M + 1$ nodes. Suppose $C2_k$, where $k \in \{1, \cdots, M-1\}$, is the value that maximizes the correlation function.

.3. Repeat the above step until the length of each subinterval is smaller than some satisfactory number, for example 0.1 or 0.05. When this condition is satisfied, the node which gives maximum value of the objective function is considered the practically maximizer of the locally weighted rank correlation function, and is used as the estimator of $c(\cdot)$.

To illustrate whether this searching procedure is valid, in simulation studies, at any point $U = u$ the initial interval is specified as $[true\ c(\cdot) - 10, true\ c(\cdot) + 10]$. Firstly, this interval is divided into 200 subintervals, which involves subinterval length equals 0.1 and 201 calculations of the locally weighted rank correlation. Secondly, the efficient fast searching procedure is conducted, with $M = 10$ and the final length of the subinterval is set to be smaller than 0.1. This would involve in total only around 40 calculations of the locally weighted rank correlation. Simulation results indicate that the time-saving efficient searching procedure is valid. To improve the validity of this searching algorithm, one can use larger updating searching intervals. For instance, instead of updating the maximizer between $L_n = Cn_{k-1}$ and $R_n = Cn_{k+1}$, one could update the maximizer between $L_1 = C1_{k-S}$ and

$R_1 = C1_{k+S}$, where S is an integer larger than 1.

3.4.4 Estimation of the unknown link $g(\cdot)$

According to the results above, the local log-likelihood function of $g(\cdot)$ can be conducted easily. To find the maximizer of the log-likelihood function, initial values of $g(\cdot)$ and $g^{'}(\cdot)$ are required. Let $t_i = \mathbf{X}^T \hat{\boldsymbol{\beta}}(U_i)$ for $i = 1, \cdots, n$. For simplicity, treat y_i $i = 1, \cdot, n$ are natural observations from model

$$y = g(t) + \epsilon,$$

where $\epsilon \sim N(0, \delta)$. Apply a standard local linear modelling and obtain the initial estimates as $\tilde{g}(t)$ and $\tilde{g}^{'}(t)$.

When initial estimates are obtained, consider the object function

$$\ell(d, q) = \frac{1}{n} C_1(\boldsymbol{\phi}) \sum_{i=1}^{n} f\left(d + q(t_i - t),\ y_i\right) K_{h_1}(t_i - t) + \sum_{i=1}^{n} C_2(y_i, \boldsymbol{\phi}) K_{h_1}(t_i - t). \tag{3.23}$$

Its first and second derivatives are given by

$$\begin{aligned} \dot{\ell}(d) &= \frac{1}{n} C_1(\boldsymbol{\phi}) \sum_{i=1}^{n} \dot{f}\left(d + q(t_i - t),\ y_i\right) K_{h_1}(t_i - t), \\ \dot{\ell}(q) &= \frac{1}{n} C_1(\boldsymbol{\phi}) \sum_{i=1}^{n} \dot{f}\left(d + q(t_i - t),\ y_i\right) (t_i - t) K_{h_1}(t_i - t), \end{aligned}$$

and

$$\begin{aligned}
\ddot{\ell}(d) &= \frac{1}{n}C_1(\boldsymbol{\phi})\sum_{i=1}^{n}\ddot{f}\left(d+q(t_i-t),\ y_i\right)K_{h_1}(t_i-t),\\
\ddot{\ell}(dq) &= \frac{1}{n}C_1(\boldsymbol{\phi})\sum_{i=1}^{n}\ddot{f}\left(d+q(t_i-t),\ y_i\right)(t_i-t)K_{h_1}(t_i-t),\\
\ddot{\ell}(q) &= \frac{1}{n}C_1(\boldsymbol{\phi})\sum_{i=1}^{n}\ddot{f}\left(d+q(t_i-t),\ y_i\right)(t_i-t)^2K_{h_1}(t_i-t).
\end{aligned}$$

Denote by $\dot{\ell} = \begin{pmatrix} \dot{\ell}(d) \\ \dot{\ell}(q) \end{pmatrix}$ and $\ddot{\ell} = \begin{pmatrix} \ddot{\ell}(d) & \ddot{\ell}(dq) \\ \ddot{\ell}(qd) & \ddot{\ell}(q) \end{pmatrix}$ the related vector and matrix form of the above derivatives (Note that $\ddot{\ell}(dq) = \ddot{\ell}(qd)$). One keeps updating d and q by

$$d_{n+1} = d_n - e_1^T\ddot{\ell}^{-1}\dot{\ell},$$

and

$$q_{n+1} = q_n - e_2^T\ddot{\ell}^{-1}\dot{\ell}.$$

iteratively until convergence gives the Newton-Raphson estimates of d and e.

Denote by $\tilde{g}_0(t)$ the maximizer of the log-likelihood function. The longest non-decreasing subsequence of $\tilde{g}_0(t)$, say $\tilde{g}(t_l)$, is utilized as the initial estimator of $g(\cdot)$. Treating that $\tilde{g}(t_{li})$, $i = 1, \cdots, n$ is a set of realization from an univariate non-parametric regression model

$$\eta = g(t) + \epsilon.$$

A standard local linear modelling gives the estimator of $g(t)$ as

$$\hat{g}(t) = {e_1}^{\mathrm{T}} \left(\mathbf{T}^{\mathrm{T}} W \mathbf{T}\right)^{-1} \mathbf{T}^{\mathrm{T}} W \boldsymbol{\eta},$$

where

$$\mathbf{T} = \begin{pmatrix} 1 & t_{l1} - t \\ \vdots & \vdots \\ 1 & t_{lT} - t \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} \tilde{g}(t_1) \\ \vdots \\ \tilde{g}(t_T) \end{pmatrix},$$

and

$$W = \operatorname{diag}\left(K_{h_{l2}}(t_{l1} - t),\ \cdots,\ K_{h_{l2}}(t_{lT} - t)\right).$$

4 Simulation Studies: Poisson Regression

In statistics, Poisson regression is a form of regression analysis frequently used to model count data. It assumes that the response variable y is from Poisson distribution. In practice, it is frequently assumed that the logarithm of the mean regression function of y can be modelled by a linear combination of unknown parameters. In this section, the thesis is going to investigate the maximum rank correlation estimation method through simulation studies of Poisson Regression. Before going through simulation studies, it is necessary to recall the aims and difficulties of this study.

• **Aims:** First and foremost, of interest is to get access to the varying coefficients given that the link function is monotonic. Secondly, the unknown link function is to be estimated, therefore, its monotonicity could be verified.

• **Hurdles:** The proposed estimation method in this paper involves many hurdles to be crossed. The full estimation procedure consists of quite a few stages - the direction and the norm of the varying coefficients, and the link transformation function. In each of these stages, there are corresponding tuning parameters to be handled.

4.1 Data generation

Suppose the response variable y is from the exponential family, and it is from Poisson distribution conditional on $\{\mathbf{X}, U\}$, where $\mathbf{X} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and U are the covariates. Denote the mean regression function of y given $\{\mathbf{X}_i, U_i\}$ by $\lambda(\mathbf{X}_i, U_i)$, $i = 1, \cdots, n$. Then the conditional density of y given $\mathbf{X}_i$ and U_i could be expressed as

$$f(y = y_i|\mathbf{X}_i, U_i) = \frac{\lambda(\mathbf{X}_i, U_i)^{y_i}}{y_i!} e^{\lambda(\mathbf{X}_i, U_i)}, \quad i = 1, \cdots, n.$$

In practice, frequently applied link function is the *log* transformation, *i.e.*

$$log\,(\lambda(\mathbf{X}_i, U_i)) = \mathbf{X}_i{}^T \boldsymbol{\beta}(U_i), \quad i = 1, \cdots, n.$$

In this thesis, the conditional mean regression function $\lambda(\mathbf{X}_i, U_i)$ is assumed to be linear via an unknown monotonic transformation

$$g\,(\lambda(\mathbf{X}_i, U_i)) = \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}(U_i), \quad i = 1, \cdots, n,$$

where $\boldsymbol{\beta}(\cdot)$ is the p dimensional varying coefficient vector. Of interest is to estimate the vector of varying-coefficients $\boldsymbol{\beta}(\cdot)$, and the unknown monotonic link function $g(\cdot)$.

In the following simulation studies, data are generated from Poisson regression with small dimension of varying coefficients, and the *log* transformation is used as the underline authentic monotonic link function, for simplicity. The covariates $\mathbf{X} = (X_1,\ X_2)$, where X_1 and X_2 are *i.i.d* are drawn from standard $Normal$ distribution

$$\mathbf{X} \sim \mathbf{N}(0_p, \mathbf{1}_p),$$

and U is generated from

$$U \sim U[0, 1].$$

Values of the mean regression function $\lambda(\mathbf{X}_i, U_i)$ are then computed via the *log* transformation function given the true varying coefficients. The response variable y_i is thus drown from poisson distribution with parameter $\lambda(\mathbf{X}_i, U_i)$. Simulations with respect to the following three examples under the frame of Poisson regression are conducted.

Example 1 $\beta_1(U) = sin(2\pi U)$; $\beta_2(U) = cos(2\pi U)$.

Example 2 $\beta_1(U) = sin(3\pi U)$; $\beta_2(U) = cos(2\pi U)$.

Example 3 $\beta_1(U) = sin(\pi U) + 0.6$; $\beta_2(U) = cos(2\pi U) - 0.2$.

For each of these examples, 100 simulations with sample size $n = 200$, $n = 400$ and $n = 800$ are treated using proposed estimation techniques. Throughout this section, Epanechnikov Kernel $K(t) = 0.75(1 - t^2)_+$, which is practically considered the optimal bandwidth due to its minimax property, is used. For details about the minimax property, see Fan and Gijbels (1996). At this stage of simulation, the proposed method is proceeded with respect to constant smoothing parameters.

To make the presentation clearer, denote different maximization methods as MRC_1 with updating algorithm (3.17) for Method 1; MRC_2 with updating algorithm (3.18) for Method 1; and MRC_3 for Method 2. For Method 2, of which the special case where the first component of the varying coefficients is strictly positive, the initial estimates are provided by MRC_2. Simulations indicate that, when Method 2 is applied, the estimation accuracy is sensitive to the initial estimates. There is evidence that MRC_2 provides proper estimates of the directions of the varying coefficients. Thus, MRC_2 is used to provide initial estimates. It is not necessary to spend time for MRC_2 to converge. One or two steps of iteration in maximization is sufficient.

4.2 Selection of constant tuning parameters δ and λ for one-step estimation of $\boldsymbol{\beta}_0(\cdot)$

First and foremost, whichever method is applied to access the directions of the varying coefficients, the selection of the tuning parameters δ and λ (λ is excluded from the case when strictly positive impact of covariates is identified) is an unavoidable issue. Since, tuning parameters δ and λ are crucial for estimating the unit circle varying coefficients. Before implementation of the full MRCE method, the thesis firstly evaluates the impact of the tuning parameters δ and λ on both the bandwidth selection and the estimation performance of the unit circle varying coefficients $\beta_{01}(\cdot)$ and $\beta_{02}(\cdot)$.

The estimation for a general generalized varying coefficient model involves both δ and λ. This thesis suggests to use some small positive number for λ, and δ is defined to be proportional to the range of the denominator within the smoothing approximation. For instance, at any point u, the objective function is constructed as

$$\ell(\mathbf{a}) = \frac{1}{n(n-1)} \sum_{i \neq j} I(y_i > y_j) \Phi\left(\frac{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{a}}{\delta}\right) K_{h_1}(U_i - u) K_{h_1}(U_j - u) - \lambda \mathbf{a}^{\mathrm{T}}\mathbf{a}.$$

Denote the range of $(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{a}$ by $\{R_{min}, R_{max}\}$. Then,

$$\delta = M(R_{max} - R_{min})$$

is used as the tuning parameter in the maximization, where $0 < M \leq 1$ is a positive real number. The tuning parameter δ can be defined similarly for Method 2.

- **Investigation of δ and λ for Method 1**

Firstly, the impact of the combination of δ and λ is investigated with respect to all the three examples with sample size $n = 400$, respectively, regardless of whether there exists strictly positive coefficient. Since the second stage of the two-step estimation does not involve tuning parameters δ and λ, at this stage, only the one-step estimation method for estimating the directions of the varying coefficients is considered. The investigation is designed as the following. By allowing λ and δ to vary in intervals $[0.01, 1.]$ and $[0.1, 1.]$, respectively, the mean integrated squared errors (MISE) of estimating the direction of the varying coefficients are calculated. To give a demonstration of overall estimation performance, average of the MISEs for each of the p estimators is calculated. The bandwidth used is selected as the one which minimizes the MISE of estimating $\boldsymbol{\beta}_0(\cdot)$. Both the bandwidths and the resulting MISEs of $\hat{\beta}_0(\cdot)$ are presented in Table 1.

Example 1

MRC_1	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.0997(0.104)	0.0104(0.104)	0.0111(0.104)	0.0117(0.104)
$\delta = 0.5$	0.0099(0.104)	0.01(0.104)	0.01(0.104)	0.01(0.104)
$\delta = 1.$	0.01(0.104)	0.01(0.104)	0.01(0.104)	0.01(0.104)
MRC_2	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.01(0.104)	0.0106(0.104)	0.0127(0.104)	0.0145(0.104)
$\delta = 0.5$	0.0095(0.104)	0.0102(0.104)	0.0103(0.104)	0.0103(0.104)
$\delta = 1.$	0.0093(0.104)	0.0095(0.104)	0.0097(0.104)	0.0098(0.104)

Example 2

MRC_1	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.0151(0.086)	0.0158(0.086)	0.0163(0.086)	0.0171(0.086)
$\delta = 0.5$	0.0152(0.086)	0.0152(0.086)	0.0152(0.086)	0.0152(0.086)
$\delta = 1.$	0.0153(0.086)	0.0153(0.086)	0.0153(0.086)	0.0153(0.086)
MRC_2	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1.$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.0151(0.086)	0.016(0.086)	0.0171(0.086)	0.0186(0.086)
$\delta = 0.5$	0.0151(0.086)	0.0152(0.086)	0.0152(0.086)	0.0155(0.086)
$\delta = 1.$	0.0153(0.086)	0.0154(0.086)	0.0154(0.086)	0.0156(0.086)

Example 3

MRC_1	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.0071(0.124)	0.0077(0.124)	0.0093(0.124)	0.0127(0.124)
$\delta = 0.5$	0.0076(0.124)	0.0076(0.124)	0.0076(0.124)	0.0076(0.124)
$\delta = 1.$	0.0079(0.124)	0.0079(0.124)	0.0079(0.124)	0.0079(0.124)
MRC_2	$\lambda = 0.01$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
$\delta = 0.1$	-	-	-	-
$\delta = 0.2$	0.0072(0.124)	0.0079(0.124)	0.0105(0.124)	0.0156(0.124)
$\delta = 0.5$	0.0076(0.124)	0.0077(0.124)	0.0077(0.124)	0.0077(0.124)
$\delta = 1.$	0.0080(0.124)	0.0081(0.124)	0.0081(0.124)	0.0081(0.124)

Table 1: Tuning parameters investigation for Method 1.

'-' means none-convergent estimates.

Table 1 demonstrates that the estimation performance of MRC_1 and MRC_2 are comparable. The selection of the tuning parameter δ and λ is crucial to our MRC estimation in terms of estimation accuracy. Although theoretically, δ should tend to 0, when δ is too small, the Newton-Raphson maximization algorithm does not converge at all. This means that δ is a more important factor comparing with λ in terms of computational convergence. When δ is sufficiently large, the estimation performance tends to be stable. While the tuning parameter λ is introduced to increase the speed of computational convergence, it is noticed that the estimation performance is not sensitive to λ. As long as δ ensures the computational convergence, smaller λ is preferable.

An additional phenomenon is that the practical, optimal bandwidth is not sensitive to the selection of δ and λ. For different combinations of δ and λ, although the computation time and estimation performance vary, the optimal bandwidth is stable. Therefore, when one tries to let the data itself select tuning parameters for estimating the directions, the thesis suggests that δ and λ would not have much impact on the bandwidth selection. Further, the thesis concludes that a proper combination of constant tuning parameters δ and λ is sufficient. And when trying to let the data itself select tuning parameters, data-driven δ and λ are not taken into consideration. In further simulations, it is recommended to use constant tuning parameter $\delta = 0.2$ and $\lambda = 0.1$ which will give proper estimates for the directions of the varying coefficients.

- **Investigation of δ for Method 2**

The impact of δ for the case where positive coefficients are identified needs to be identified as well. 100 simulations with respect to Example 3 with sample size $n = 400$ is conducted. For simplicity, at this stage, only the one-Step estimation method for estimating the directions of the varying coefficients is considered, since the second stage of the two-step estimation method does not involve tuning parameter δ. It has been noticed that although one has one less tuning parameter λ to consider, the selection of δ is not becoming any easier. Simulations suggest that a small tuning parameter δ is preferable.

Let δ vary in $[0.01, 0.05]$. The mean integrated squared errors (MISEs) of estimators, $\beta_{0j}(\cdot)$, $j = 1, 2$ are calculated. Again, to give a demonstration of overall estimation performance. Average of the MISEs for each of the p estimators is calculated. The bandwidth used is selected as the one which minimizes the MISE for estimating $\beta_0(\cdot)$. Both the bandwidth and the resulting MISE of $\hat{\beta}_0(\cdot)$ are presented in Table 2.

	$\delta = 0.02$	$\delta = 0.03$	$\delta = 0.04$	$\delta = 0.05$
MRC_3	0.0051(0.179)	0.0027(0.179)	0.0029(0.179)	0.003(0.179)

Table 2: Tuning parameter investigation for Method 2.

When trying to estimate the directions with this method, it has to be emphasized that at least one component of the varying coefficient functions has to be strictly positive. Table 2 indicates that for the purpose of estimation accuracy, small tuning parameter δ is preferable. Compare with Method 1, the MISE of the estimator of the directions is reduced from 0.007 in Method 1 to 0.003 in Method 2, with the practical bandwidth increased from 0.124 to 0.179. That is to say the estimation performance of Method 2 would be significantly better once known that there is positive impact of at least one covariate. In addition, it is also witnessed that the bandwidth selection procedure is not sensitive to δ. Thus, again, it is suggested to use constant δ. However, when δ is too small, the Newton-Raphson maximization algorithm faces difficulty in computational convergence. To ensure that **Method 2** retains stable and proper estimates for the directions of the varying coefficients, the thesis uses constant tuning parameter $\delta = 0.05$ throughout this paper.

4.3 Estimating the directions of the varying coefficients: a comparison between one-step and two-step method

The thesis has proposed two methods for estimating the directions of varying coefficients. The one-step method estimates the directions all together; while the two-step method serves the same purpose in two stages. Two-step estimation is a crucial method to deal with cases that the directions of the varying coefficient functions may possess different levels of smoothness. In such cases, estimation of the directions of the varying coefficient functions require different smoothing parameters.

Theoretically, when it is known in advance that the directions of the different varying coefficients enjoy approximately the same level of smoothness, a one-step estimation method is sufficiently good and ready to be applied for providing estimates for the directions of the varying coefficients. However, the situation becomes more complicated when this assumption does not hold. Since estimating smoother functions requires larger bandwidth, a one-step estimation in fact uses smaller bandwidth which suffers from the variance site. In this case, the two-step estimation method is needed for rescue.

Two-step method is initially designed for smoother functions. The idea of two-step estimation is that it uses a smaller bandwidth in the first stage to make the bias negligible and in the second stage let the local modelling to control the variance site. For rougher functions, the two-step estimation would at least perform as good as the one-step estimation method. That is to say, it is not needed to worry about how to identify which functional direction of varying coefficient is smoother or rougher than others. Therefore, when two-step method is used, it is applied to all functional directions of the varying coefficients.

For Example 1 to Example 3, the directions of the varying coefficients are estimated with both one-step and two-step methods. The bandwidths are selected by the minimization of corresponding MISEs of estimators. Let the MISEs be an indicator of estimation performance. Table 3 to 3 present the estimation performance for all three examples. In terms of the two-step estimation method, the first stage bandwidth is set as $h_{20} = 0.5h_1$, where h_1 is the optimal bandwidth used for the one step estimation.

- **Example 1** $\beta_1(U) = sin(2\pi U)$; $\beta_2(U) = cos(2\pi U)$.

In Example 1, the two varying coefficients have the same order of smoothness. As the norm of the varying coefficient vector is 1, corresponding functional directions of the varying coefficients also have the same order of smoothness. Although, a one-step estimation approach shall be sufficient, the two-step method is as well implemented and compared.

Sample size	n=200		n=400		n=800	
Method	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_{01}(\cdot)\}$	0.022	0.013	0.012	0.007	0.008	0.004
$MISE\{\hat{\beta}_{02}(\cdot)\}$	0.016	0.008	0.009	0.004	0.005	0.002

Table 3: Example 1: Mean integrated squared errors of $\hat{\beta}_{0j}(\cdot), j = 1, 2$

Table 3 shows that as sample size increases, both one-step and two-step methods steadily improve the estimation accuracy of the directions of the varying coefficients. Compare with one-step method, the increments in terms of estimation performance for both $\beta_{01}(\cdot)$ and $\beta_{02}(\cdot)$ are significant. Take the second component $\beta_{02}(\cdot)$ as a demonstration. The mean integrated squared error of $\hat{\beta}_{02}(\cdot)$ is halved in simulations of all three sample sizes.

Thus, for the case when varying coefficients are of similar smoothness, it is ideal to use two-step estimation method for generating estimates of the directions of the varying coefficients. To give a visual demonstration of the estimation, Figure 1 plots the estimates of the directions of the varying coefficients for different sample sizes that have the medium estimation errors among 100 simulations.

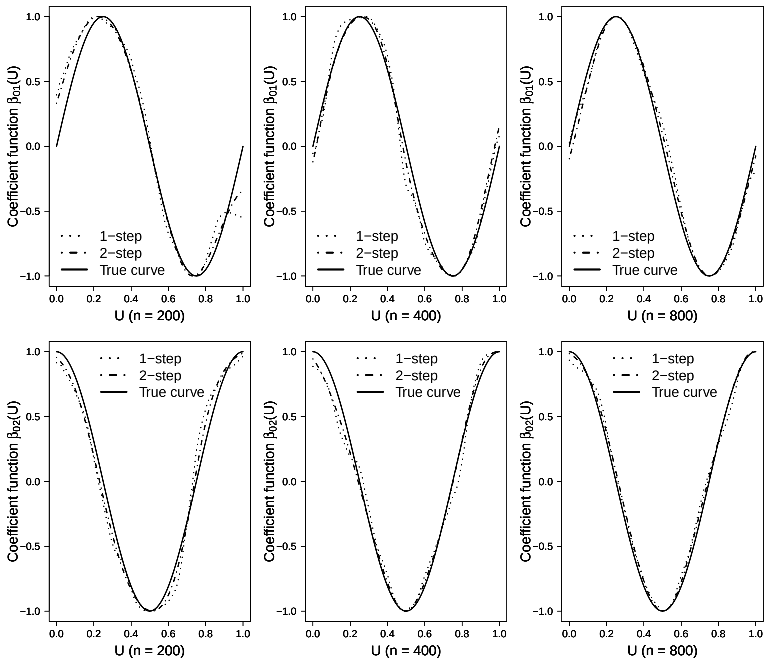


Figure 1: Direction estimation for Example 1.
The plots depict $\hat{\beta}_{0j}(\cdot)$, $j = 1, 2$, that have medium integrated squared errors among 100 simulations.

- **Example 2** $\beta_1(U) = sin(3\pi U)$; $\beta_2(U) = cos(2\pi U)$.

In Example 2, the second varying coefficient is smoother than the first one. The thesis as well estimates the directions of the two varying coefficients with both one-step and two-step methods. Estimation results are depicted in Table 4 and Figure 2.

Sample size	n=200		n=400		n=800	
Method	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_{01}(\cdot)\}$	0.029	0.024	0.019	0.016	0.01	0.009
$MISE\{\hat{\beta}_{02}(\cdot)\}$	0.022	0.018	0.014	0.012	0.006	0.005

Table 4: Example 2: Mean integrated squared errors of $\hat{\beta}_{0j}(\cdot), j = 1, 2$

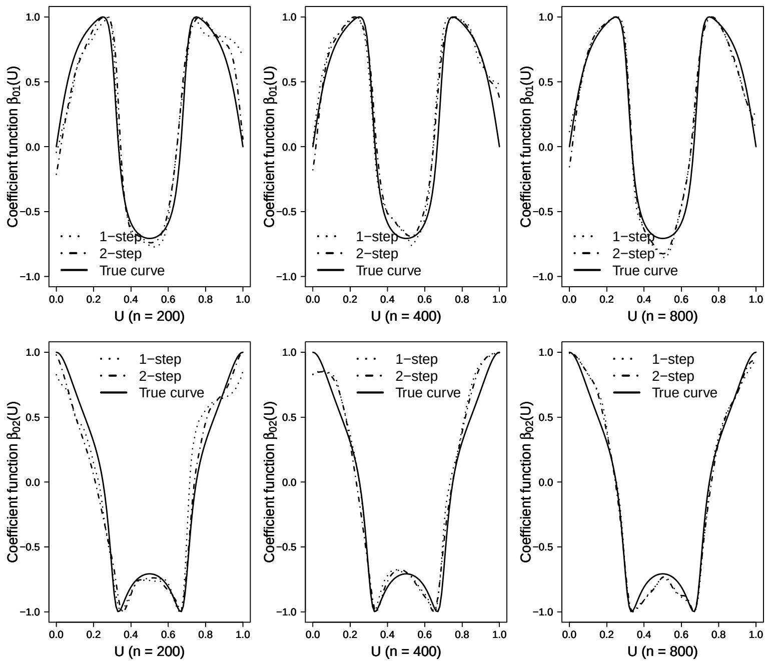


Figure 2: Direction estimation for Example 2.

The plots depict $\hat{\beta}_{0j}(\cdot)$, $j = 1, 2$, that have medium integrated squared errors among 100 simulations.

Table 4 witnesses stable reduction of estimation error as sample size increases from 200 to 800. The two-step estimation method gives better estimates of the directions, especially when the sample size is small. Despite that the increment of estimation accuracy reduces as the sample size increases, the two-step method is as well a preferable choice for obtaining estimates of $\boldsymbol{\beta}_0(\cdot)$. Again the estimates of the directions of the varying coefficients with medium estimation errors are plotted out to give a visual insight. Both one-step and two-step methods capture the functional directions properly.

- **Example 3** $\beta_1(U) = sin(\pi U) + 0.6$; $\beta_2(U) = cos(2\pi U) - 0.2$.

In Example 3, the first component of the varying coefficients is strictly positive. We first compare the estimation performance between Method 1 and Method 2 with one-step method for simplicity.

Sample size	n=200		n=400		n=800	
Method	1	2	1	2	1	2
$MISE\{\hat{\beta}_{01}(\cdot)\}$	0.0067	0.0034	0.0039	0.0026	0.0024	0.0015
$MISE\{\hat{\beta}_{02}(\cdot)\}$	0.0209	0.0109	0.0118	0.0053	0.0066	0.0034

Table 5: Example 3: A comparison between Method 1 and 2.

It is seen in Table 5 that Method 2 is dramatically more accurate in estimating the directions of the varying coefficients than Method 1. The mean integrated squared errors returned by Method 2 is nearly half the values of that given by Method 1. Therefore, when it is confident that some covariate impacts on the response variable positively, Method 2 is preferable than Method 1.

For Example 3, the one-step and two-step methods using Method 2 are compared and demonstrated in Table 6.

Sample size	n=200		n=400		n=800	
Method 2	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_{01}(\cdot)\}$	0.0034	0.0023	0.0026	0.0017	0.0015	0.0004
$MISE\{\hat{\beta}_{02}(\cdot)\}$	0.0109	0.008	0.0053	0.0023	0.0034	0.0013

Table 6: Example 3: Mean integrated squared errors of $\hat{\beta}_{0j}(\cdot), j = 1, 2$

Both one-step and two-step method approaches the directions of the varying coefficients reasonably, while comparing with the one-step method, Table 6 again witnesses dramatic reduction in estimation errors when the two-step method is used. Such improvement may be visually clear in Figure 3, which plots the estimates of the directions with medium integrated squared errors.

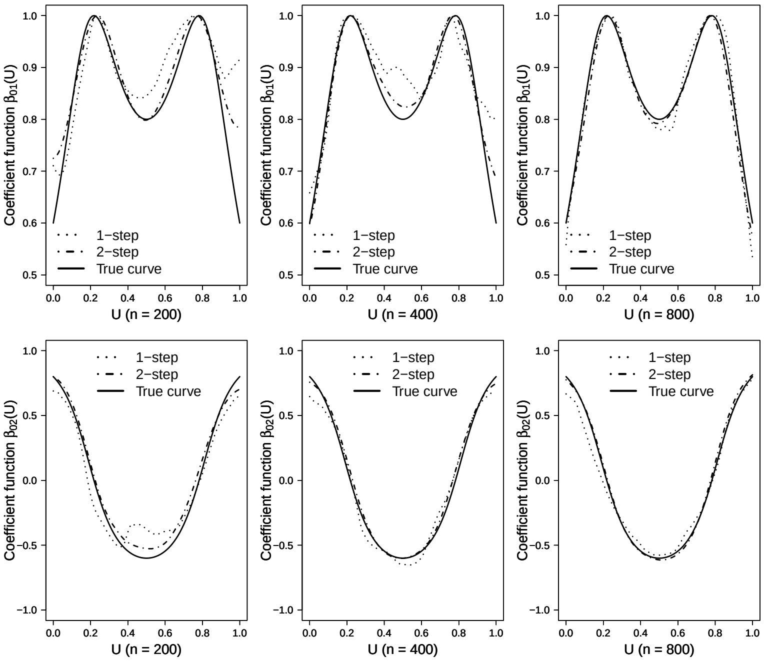


Figure 3: Direction estimation for Example 3.
The plots depict $\hat{\beta}_{0j}(\cdot)$, $j = 1, 2$, that have medium integrated squared errors among 100 simulations.

Simulations of all three examples suggest that the maximum rank correlation method works well with the approximation of the directions of the varying coefficients. Both one-step and two-step methods are satisfactory approaches. While the two step estimation method improves dramatically to the one-step estimation method, even when the varying coefficients possess the same smoothness. This finding supports that, in practice, it is confident to use the two step method for the estimation of the directions of the varying coefficients without thinking about the identification of the difference of smoothness of the functional directions.

4.4 Estimation of the varying coefficients

With the estimation of the directions of the varying coefficients being simulated, in this section, the task is to investigate the maximum rank correlation method which provides estimates for the unknown norm and hence the varying coefficients. Unlike the estimation of the directions of the varying coefficients, which involves the maximization of an approximated smoothing rank correlation function, the norm estimation is univariate and is conducted with the idea of grid regression. To control the expense of computational cost, the fast searching algorithm proposed is recommended.

Standard estimation of the varying coefficients

- **Example 1** $\beta_1(U) = sin(2\pi U)$; $\beta_2(U) = cos(2\pi U)$.

In Example 1, $\|\boldsymbol{\beta}(u)\| \equiv 1$ for any $U = u$. The target true maximiser of $c(\cdot)$ is $c(u) = 0$ at any $U = u$. When applying the grid regression, the maximiser of $c(u)$ at any $U = u$ is searched for between -20 and 20. The bandwidth h_n is chosen to be the one that minimizes the average value of the mean integrated squared errors of $\hat{\beta}_j(\cdot)$, $j = 1, 2$. In Table 7, the mean integrated squared errors of the norm and corresponding varying coefficients are presented. Estimation of the norm involves estimates of the directions of the varying coefficients. To identify the impact of direction estimation on the estimation of the norm, the directions are estimated with both one-step and two-step methods respectively.

Sample size	n=200		n=400		n=800	
Method	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{N}(\cdot)\}$	0.0023	0.0016	0.0011	0.0008	0.0003	0.0003
$MISE\{\hat{\beta}_1(\cdot)\}$	0.023	0.0135	0.0124	0.0071	0.008	0.004
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0167	0.009	0.0098	0.0046	0.0051	0.0023

Table 7: Example 1: the mean integrated squared errors.

For Example 1, the estimation method for the directions does not challenge on the bandwidth selection for h_n. For both one-step and two-step methods, the practical bandwidths h_n which minimize the mean integrated squared errors of $\hat{\boldsymbol{\beta}}(\cdot)$ are identical. The two-step estimation method provides better estimates of the directions of the varying coefficients. When two-step estimates $\hat{\boldsymbol{\beta}}_0(U_i)$, $i = 1, \cdots, n$, are used for norm estimation, estimates of the norm given by the grid regression are closer to the true norm for samples sized $n = 200$ and $n = 400$. When the sample size increases to 800, although the mean integrated squared errors of $\hat{\boldsymbol{\beta}}_0(\cdot)$ returned by one-step method is still twice that of two-step method, the norm is estimated as accurately as that of the two-step method. The advantage of two-step estimation method aggregates and becomes more significant when the estimator $\hat{\boldsymbol{\beta}}(\cdot)$ is composed. The mean integrated squared errors of $\hat{\boldsymbol{\beta}}(\cdot)$ obtained via two-step estimation method is dramatically smaller than that of the one-step method. Thus, the two-step method is not only a crucial tactic for direction estimation, it is promising in all stages of the estimation of the varying coefficients.

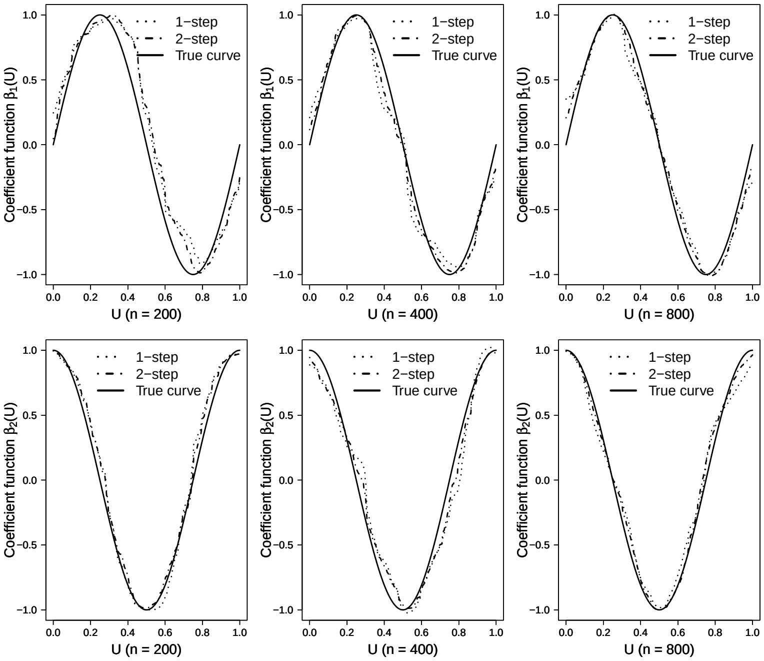


Figure 4: Varying coefficient estimation for Example 1.
The plots depict $\hat{\beta}_j(\cdot)$, $j = 1, 2$, that have medium integrated squared errors among 100 simulations.

Figure 4 plots the estimated curves of the varying coefficients with both one-step and two-step methods against the true functional varying coefficients. Both one-step and two-step methods approach the varying coefficients properly, while the two-step method is considered preferable.

- **Example 2** $\quad \beta_1(U) = sin(3\pi U); \qquad \beta_2(U) = cos(2\pi U).$

For Example 2, the advantage of two-step estimation of the directions of the varying coefficients is less dramatic than that of the one-step method. Therefore, it is anticipated to witness increment in the estimation of the norm and varying coefficients as significant as in Example 1. The target true $c(\cdot)$

is calculated at any $U = u$. Denote these true values as $c_t(U_i)$, $i = 1, \cdots, n$. When applying the grid regression at any $U = u$, the estimator of $c(u)$ is seek between $c_t(u) - 20$ and $c_t(u) + 20$. The bandwidth h_n is chosen to be the one that minimizes the average value of the mean integrated squared errors of $\hat{\beta}_j$, $j = 1, 2$. The mean integrated squared errors of the norm and corresponding varying coefficients are presented in Table 8.

Sample size	n=200		n=400		n=800	
Method	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{N}(\cdot)\}$	0.088	0.085	0.063	0.063	0.034	0.032
$MISE\{\hat{\beta}_1(\cdot)\}$	0.06	0.059	0.048	0.047	0.027	0.026
$MISE\{\hat{\beta}_2(\cdot)\}$	0.071	0.061	0.04	0.038	0.021	0.019

Table 8: Example 2: the mean integrated squared errors.

In Example 2, again, the practical bandwidth h_n which minimizes the mean integrated squared errors of $\hat{\boldsymbol{\beta}}(\cdot)$ is identical for both one-step and two-step methods, indicating a weak impact from direction estimation onto bandwidth selection of norm estimation. The increment in norm estimation for two-step method is negligible for Example 2. However, the two-step method is still doing better in all stages of the estimation procedure. Two-step method for the rougher function $\beta_1(\cdot)$ is at least not worse than that of one-step method. While for the smoother function $\beta_2(\cdot)$, the reduction in estimation errors is more significant. Thus, the two-step method is still crucial for obtaining estimates of the varying coefficients for Example 2.

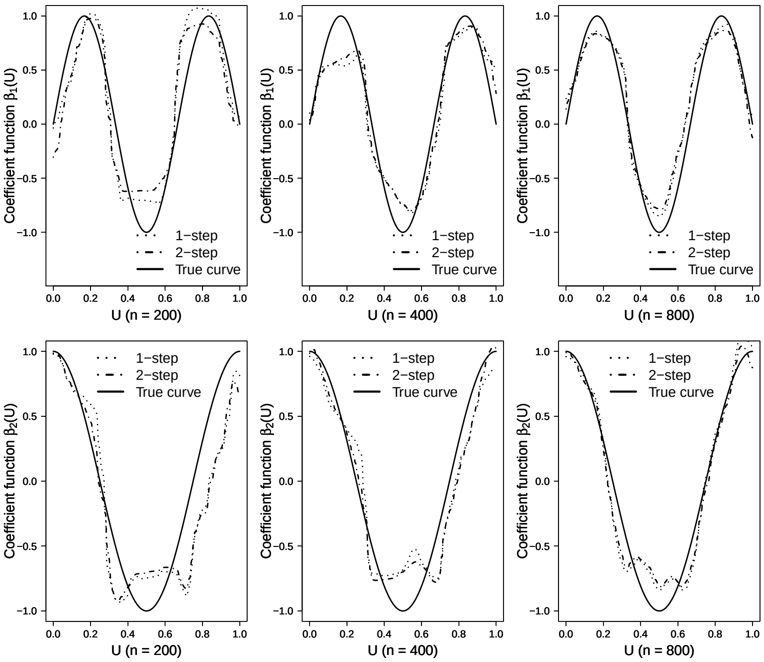


Figure 5: Varying coefficient estimation for Example 2.
The plots depict $\hat{\beta}_j(\cdot)$, $j = 1, 2$, that have medium integrated squared errors among 100 simulations.

The varying coefficients are shown in Figure 5 as well for visual presentation. As the varying coefficients are more complicated than that of Example 1, it is found that small sample size may be an obstacle in this case.

Improved estimation of the varying coefficients

The thesis has implemented the idea of grid regression in the estimation of the unknown norm. However, the estimation of the norm could be very difficult when bandwidth h_n is small. This is because, the estimation of the norm depends on the estimation of its derivatives, which is statistically difficult to estimate. The difficulty aggregates when estimation at the left boundary of U is inaccurate and unstable. Recall that to deal with the identifiability issue, it is set that $\|\boldsymbol{\beta}(0)\| = 1$. When one has the estimator $\hat{c}(\cdot)$, the estimator of the unknown norm is constructed as

$$\hat{N}(u) = \exp\left\{\int_0^u \hat{c}(u)du\right\},$$

and the estimator of the varying coefficient vector is obtained via

$$\hat{\boldsymbol{\beta}}(u) = \hat{N}(u)\hat{\boldsymbol{\beta}}_0(u).$$

There is a potential threat here. As the integration starts from the left boundary $U = 0$, when the estimates of $c(\cdot)$ is extremely poor near $U = 0$, the poorness would impact on the estimates of the norm at all following points. To reduce the risk of poor estimates of $c(u)$ near the left boundary of U, what the thesis suggests is to reconsider the identifiability condition $\|\boldsymbol{\beta}(0)\| = 1$.

The condition, $\|\boldsymbol{\beta}(0)\| = 1$, is set to avoid the identifiability issue frequently confronted in the maximum rank correlation estimation. However, it constrains on the estimation of the unknown norm, which integrates $c(\cdot)$ from $U = 0$. Thus, it may be wiser to set $\|\boldsymbol{\beta}(u)\|$ is known for another u, where stable and proper estimates of $c(\cdot)$ near $U = u$ are confidently expected. The validity of this idea is tested for Example 1 and 2 by assuming $\|\boldsymbol{\beta}(U = 0.2)\|$. When one has estimates of $\hat{c}(\cdot)$, the estimator of the unknown norm is constructed as

$$\hat{N}(u) = \exp\left\{\int_{0.2}^u \hat{c}(u)du\right\}\|\boldsymbol{\beta}(0.2)\|,$$

and the estimator of the varying coefficient vector is obtained via

$$\hat{\boldsymbol{\beta}}(u) = \hat{N}(u)\hat{\boldsymbol{\beta}}_0(u).$$

For Example 1 and 2, the thesis tests and compares this idea with the standard estimation procedure implemented in previous section.

- **Example 1** $\beta_1(U) = sin(2\pi U)$; $\beta_2(U) = cos(2\pi U)$.

In Example 1, $\|\boldsymbol{\beta}(u)\| \equiv 1$ at any $U = u$. Thus, the target true maximiser $c(\cdot)$ is $c(u) = 0$ at any $U = u$. We search for the maximiser of $c(u)$ at any $U = u$ between -20 and 20. The bandwidth h_n is chosen to be the one that minimizes the average value of the mean integrated squared errors of $\hat{\beta}_j(\cdot)$, $j = 1, 2$. The directions of the varying coefficients are estimated with the two-step method.

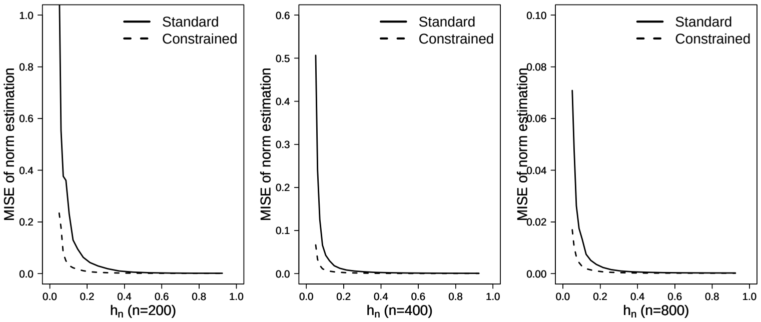


Figure 6: Example 1: MISE against bandwidth h_n.
The solid lines are the MISEs of $\hat{N}(\cdot)$ against bandwidth h_n given $\|\boldsymbol{\beta}(0)\| = 1$; The dashed lines are the MISEs of $\hat{N}(\cdot)$ against bandwidth h_n given $\|\boldsymbol{\beta}(U = 0.2)\|$.

Figure 6 compares the mean integrated squared error of $\hat{N}(\cdot)$ for different sample sizes. One significant feature is that by assuming $\|\boldsymbol{\beta}(U = 0.2)\|$ is known, comparing with setting $\|\boldsymbol{\beta}(0)\| = 1$, the mean integrated squared error of $\hat{N}(\cdot)$ is dramatically reduced for small h_n. $\hat{c}(\cdot)$ is not impacted by the setting of the identifiability condition. What makes the difference in the estimation of the norm is the improvement in the integration of $\hat{c}(\cdot)$. When h_n is small, $\hat{c}(u)$ can be extremely poor for small u. The poorness of $\hat{c}(\cdot)$ at small $U = u$ invalidates reasonable $\hat{c}(\cdot)$ at larger $U = u$ and deteriorates the estimation of $N(\cdot)$ continuously during the integration of $\hat{c}(\cdot)$. By assuming $\|\boldsymbol{\beta}(U = 0.2)\|$ is known, the integration of $\hat{c}(\cdot)$ starts from $U = 0.2$, a relatively safe location where the estimates $\hat{c}(\cdot)$ are anticipated to be more accurate than those at smaller u. The integration of $\hat{c}(\cdot)$ under this identifiability condition gets rid of potential poor estimates of $\hat{c}(\cdot)$ for small u and results in dramatic improvement in the estimation of the norm and hence the varying coefficients, especially for small h_n.

Sample size	n=200		n=400		n=800	
Method	a	b	a	b	a	b
$MISE\{\hat{N}(\cdot)\}$	0.0016	0.0011	0.0008	0.0005	0.0003	0.0002
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0135	0.0131	0.0071	0.0067	0.004	0.0039
$MISE\{\hat{\beta}_2(\cdot)\}$	0.009	0.0088	0.0046	0.0046	0.0023	0.0023

Table 9: Example 1: comparison between standard and improved estimation. 'a' denotes standard estimation, and 'b' denotes the improved estimation.

Table 9 compare the estimation performance of the varying coefficients with two different identifiability assumptions - the standard assumption $\|\boldsymbol{\beta}(0)\| = 1$ and the modified assumption $\|\boldsymbol{\beta}(0.2)\|$ is given. The mean integrated squared errors suggest that the modified identifiability condition does not give significant improvement in the estimation of the varying coefficients. The advantage of the modified identifiability condition is useful mainly for small bandwidth h_n. Up to this stage, the bandwidths are chosen by minimization of mean integrated squared errors. When it comes to date-driven bandwidth selection, the modified identifiability condition would have more significant advantages than the standard condition.

- **Example 2** $\beta_1(U) = sin(3\pi U);$ $\beta_2(U) = cos(2\pi U).$

The target true $c(\cdot)$ for Example 2 is calculated at any $U = u$. Denote these true values as $c_t(U_i)$, $i = 1, \cdots, n$. When applying the grid regression at any $U = u$, search for the maximiser of $c(u)$ between $c_t(u) - 20$ and $c_t(u) + 20$. The bandwidth h_n is chosen to be the one that minimizes the average value of the mean integrated squared errors of $\hat{\beta}_j$, $j = 1, 2$. The directions of the varying coefficients are estimated with the two-step method.

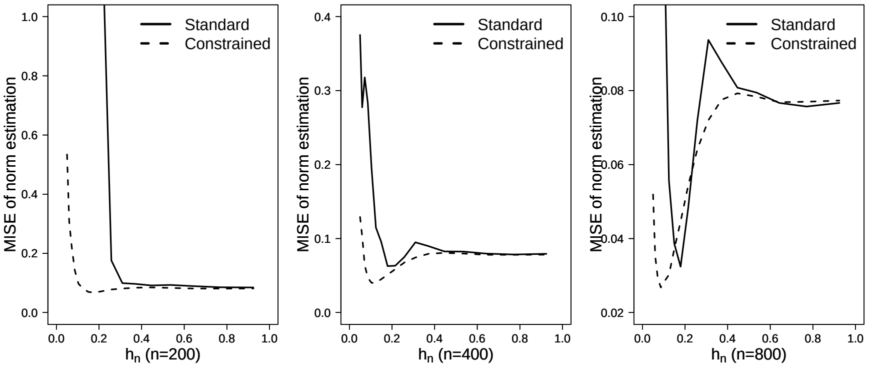


Figure 7: Example 2: MISE against bandwidth h_n.
The solid lines are the MISEs of $\hat{N}(\cdot)$ against bandwidth h_n given $\|\boldsymbol{\beta}(0)\| = 1$; The dashed lines are the MISEs of $\hat{N}(\cdot)$ against bandwidth h_n given $\|\boldsymbol{\beta}(U = 0.2)\|$.

Figure 7 has witnessed improvement in estimation of the norm, especially when the bandwidth h_n is small. The practical, optimal bandwidth h_n under the modified identifiability condition is smaller than that of the standard condition for all three sample sizes. This is in line with the fact that when h_n is small under the standard identifiability condition, the estimates of $c(\cdot)$ when u is small could be very poor, causing deterioration in the estimation of the norm, thus larger h_n is preferable. With the modified identifiability condition, the impact of the estimates of $c(\cdot)$ when u is small is reduced since the integration of $\hat{c}(\cdot)$ that provides estimates of the norm starts from larger u around which $c(\cdot)$ is better approached.

Sample size	n=200		n=400		n=800	
Method	a	b	a	b	a	b
$MISE\{\hat{N}(\cdot)\}$	0.085	0.067	0.063	0.04	0.032	0.027
$MISE\{\hat{\beta}_1(\cdot)\}$	0.059	0.049	0.047	0.031	0.027	0.019
$MISE\{\hat{\beta}_2(\cdot)\}$	0.061	0.055	0.038	0.034	0.019	0.019

Table 10: Example 2: comparison between standard and improved estimation. 'a' denotes standard estimation, and 'b' denotes the improved estimation.

Table 10 compares the estimation performance of the varying coefficients with two different identifiability assumptions - the standard assumption $\|\boldsymbol{\beta}(0)\| = 1$ and the modified condition $\|\boldsymbol{\beta}(0.2)\|$ is given. The modified condition surpasses the standard assumption in the estimation of the norm for all three sample sizes, while its advantage in the estimation of the varying coefficients is reduced as sample size increases. This is because when sample size is large, the estimates of $c(\cdot)$ for small u becomes accurate, thus the integration from $U = 0$ is not worse than that from inner points (U=0.2 in this simulation). We still conclude that the modified identifiability condition is more stable and safer than the standard assumption in providing estimates of the norm of the varying coefficients. Therefore, in later simulations and real data analysis, the modified identifiability condition is utilized.

4.5 Estimation of the unknown monotonic link function

With that the varying coefficients are estimated properly, the next task is to estimate the monotonic link function. The link function

$$g(\cdot) = log(\cdot)$$

is used for all three examples. For simplicity, data set used for the estimation of the unknown link function in this section is from Example 1. For 100 simulations with sample size $n = 200, 400$ and 800, the link function is estimated by the proposed method.

Standardized sum of squares of Person's residuals is considered as an indicator of goodness of estimation performance. Denote the standardized sum of squares of Person's residuals by

$$r_j^2 = \Big(\frac{y_i - \hat{m}(y_i|\mathbf{X}_i, U_i)}{\sqrt{\hat{m}(y_i|\mathbf{X}_i, U_i)}}\Big)^2, \quad i = 1, \cdots, n,$$

where $j = 1, \cdots, 100$ refers to the jth simulation. Consider r_j^2 as a function of h_l, $r_j^2(h_l)$. Bandwidth for the estimation of the link function is chosen by minimization of the following objective function

$$Score(h_l) = \frac{1}{100}\sum_{j=1}^{100} r_j^2(h_l),$$

which is the mean value of standardized sum of squares of Person's residuals over 100 simulations.

Sample size	n=200		n=400		n=800	
Method	1-step	2-step	1-step	2-step	1-step	2-step
Score	10.613	7.935	7.695	5.618	5.243	3.936

Table 11: Standardized Pearson's residual sum of squares.

Table 11 indicates the impact of proper estimation of the varying coefficients on the link estimation. For Example 1, when the two-step method is utilized for providing estimates of the directions of the varying coefficients, the varying coefficients are approached more accurately than that of one-step method. With the effort of two-step method, the goodness of fit is significantly improved.

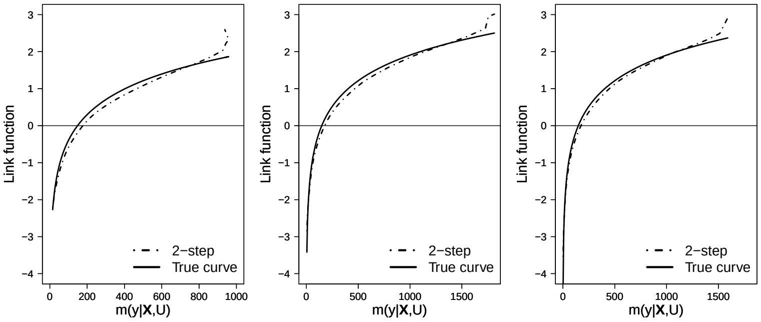


Figure 8: Link function estimation.
Figures from the left to the right are estimated link function with medium sum of squares of Pearson's residuals for sample sizes $n = 200, 400$ and 800, respectively.

5 Bandwidth Selection

Bandwidth selection is a naturally arisen question for non-parametric modelling. A bandwidth is crucial in that, it decides the trade-off between modelling bias and variance. When a smaller than optimal bandwidth is used, local modelling should provide a local estimate with small bias on one hand and large variance on the other. This is because small bandwidth includes only limited local data. In practice, proper data-driven smoothing parameters are desirable. Thus, in this section, the thesis tries to let the data itself decide the smoothing parameters.

Smoothing parameter is crucial to the maximum rank correlation method, especially when the norm is to be estimated. In this project, the estimation of the varying coefficient functions and the unknown link function involves quite a few bandwidths. We denote these bandwidths by h_1 and h_2, the one-step and two-step estimation bandwidth used for approaching the directions of the varying coefficient functions, respectively; h_n, the bandwidth applied to estimate the unknown norm; and h_l, the bandwidth employed in the estimation of the unknown link function. Bear in mind that most of the difficulties in the whole procedure of estimation lies in the endeavour to obtain accurate estimates of the unknown norm.

5.1 Data-driven constant bandwidth

When a global constant bandwidth is wanted, either the AIC criterion proposed by Cheng, Zhang and Fan (2009) or the CV criterion may be utilized as an approach for data-driven smoothing parameter selection. These two criteria are introduced briefly in this section.

AIC Criterion

Define the AIC for model (3.1) as

$$AIC(h_1, h_n, h_l) = -2 \sum_{i=1}^{n} lnL \left(\hat{g}(\mathbf{X}_i{}^T \hat{\boldsymbol{\beta}}(U_i), y_i) \right) + 2\kappa, \tag{5.1}$$

when directions are estimated using one-step method; or

$$AIC(h_{21}, \cdots, h_{2p}, h_n, h_l) = -2 \sum_{i=1}^{n} lnL \left(\hat{g}(\mathbf{X}_i{}^T \hat{\boldsymbol{\beta}}(U_i), y_i) \right) + 2\kappa, \tag{5.2}$$

when directions are estimated using two-step method; where κ is the number of parameters involved in the estimation procedure. Bandwidths that are utilized for providing initial estimates are not crucial as long as they are not stupidly selected. Thus, these bandwidths are not included in (5.1) and (5.2).

According to Fan and Gijbels (1996), when local linear approximation is used in non-parametric modelling, and when the sample size n is sufficiently large, the number of unknown parameters that an unknown function amounts to is approximately

$$h^{-1}(v_0 + v_2/\mu_2),$$

where $v_i = \int t^i K^2(t)dt$ and $\mu_i = \int t^i K(t)dt$.

Similarly, when local constant approximation is used, the number of unknown parameters that an unknown function amounts to is approximately

$$h^{-1}v_0.$$

In this project, the Epanecknicov kernel is utilized, hence $v_0 = 0.6$ and $v_0 + v_2/\mu_2 \approx 1.285714$. Therefore, the number of unknown parameters involved in our estimation procedure that counts in the AIC is

$$\kappa = 0.6h_1^{-1}p + 1.285714\left(h_n^{-1} + h_l^{-1}\right),$$

when one-step method is applied; and

$$\kappa = 1.285714\left(\sum_{j=1}^{p} h_{2j}^{-1} + h_n^{-1} + h_l^{-1}\right),$$

when two-step estimation method is used. The set of bandwidths to be used is the set that gives the minimal of the AIC score.

CV Criterion

The cross-validation approach is also straightforward. Define the CV for model (3.1) as

$$CV(h_1, h_n, h_l) = -2\sum_{i=1}^{n} lnL\left(\hat{g}^{-i}(\mathbf{X}_i{}^T\hat{\boldsymbol{\beta}}^{-i}(U_i), y_i)\right), \qquad (5.3)$$

when directions are estimated using one-step method; or

$$CV(h_{21}, \cdots, h_{2p}, h_n, h_l) = -2\sum_{i=1}^{n} lnL\left(\hat{g}^{-i}(\mathbf{X}_i{}^T\hat{\boldsymbol{\beta}}^{-i}(U_i), y_i)\right), \qquad (5.4)$$

when directions are estimated using one-step method. The estimates $\hat{\boldsymbol{\beta}}^{-i}(U_i)$, $i = 1, \cdots, n$, of the varying coefficients are provided by the data with the ith observation being removed, that is $\{\mathbf{X}_{-i}, U_{-i}, y_{-i}\}$, $i = 1, \cdots, n$. The estimates $\hat{g}^{-i}(t_i)$ of the link function are as well provided by the data with the ith observation being removed, that is $\{t_{-i}, y_{-i}\}$, $i = 1, \cdots, n$. The set of practical optimal bandwidths, h_1, h_{2_j}, $j = 1, \cdots, p$, h_n and h_l is the combination that minimizes the CV score.

Bandwidth selection algorithm

It is frustrating that we have quite a few bandwidths to select. Moreover, these bandwidths impact upon each other. If one wants to estimate the varying coefficients reasonably, initially, one has to have good estimates of both their directions and the norm. However, the accuracy of estimating the norm depends on how well the directions are approached. The bandwidth selection procedure also involves the estimates of the link function, for which reasonable estimates of the varying coefficients have to be ensured. That is to say, the estimation stages, those of the direction estimation, norm estimation and the link estimation, are inter-correlated to each other. Searching for optimal bandwidths across all candidate bandwidths is therefore time-consuming, especially when the sample size is large.

It is absolutely not easy to say there is any rule, which on one hand is computationally cheap and practically efficient on the other, to be applied. For the sake of trading off between providing reliable bandwidths and saving computational cost, we propose a bandwidth selection procedure which provides relatively reasonable bandwidths and reduces the computational cost at the same time.

- **Algorithm when one-step estimation method is used**

1. Knowing that the link function is strictly monotonic, we start bandwidth selection by setting a proper guess of $h_l = h_{l0}$. Let $h = h_1 = h_n$, using one-step estimation method, search for $\hat{h}$ that minimizes the bandwidth criteria in use.

2. Fix the bandwidth for estimating the norm as $h_n = \hat{h}$. Search for $\hat{h}_1$

that minimizes the bandwidth criteria in use. $h_1 = \hat{h}_1$ is selected as the final bandwidth that provide estimates of the directions of the varying coefficients.

3. Firstly, fix the bandwidth for estimating the directions of the varying coefficients as $h_1 = \hat{h}_1$. Then search for the bandwidth for estimating the norm again, and find $\hat{h}_n$ that minimizes the bandwidth criteria in use. $hn = \hat{h}_n$ is used as the final bandwidth that provide estimates of the norm of the varying coefficients.

4. With $h_1 = \hat{h}_1$ and $h_n = \hat{h}_n$ being fixed, search for $\hat{h}_l$ that minimizes the bandwidth criteria in use. $h_l = \hat{h}_l$ is selected to be the bandwidth for the estimation of the monotonic link function.

- **Algorithm when two-step estimation method is used**

When the two-step method is used for the estimation of the directions of the varying coefficients, the first two steps of bandwidth selection are the same as that of the algorithm with one-step estimation method.

1. Given a strictly monotonic link function, we start bandwidth selection by setting a proper guess of $h_l = h_{l0}$. Let $h = h_1 = h_n$, using one-step estimation method, search for $\hat{h}$ that minimizes the bandwidth criteria in use.

2. Fix the bandwidth for estimating the norm as $h_n = \hat{h}$. Search for $\hat{h}_1$ that minimizes the bandwidth criteria in use. Due to that two-step estimation method is not sensitive to the first state bandwidth, as long as it is not too ugly. We set $h_{20} = 0.5\hat{h}_1$ as the first stage bandwidth that provide under-smoothed estimates of the directions of the varying coefficients.

3. With the bandwidth for the estimation of the norm of the varying coefficients being fixed as $h_n = \hat{h}$, we search for a $\hat{h}_{2j}$, $j = 1, \cdots, p$, that minimize the bandwidth criteria in use. $h_{2j} = \hat{h}_{2j}$, $j = 1, \cdots, p$ are selected

to be the second stage bandwidths for the estimation of the directions of the varying coefficients.

4. With $h_{20} = 0.5\hat{h}_1$ and $h_{2j} = \hat{h}_{2j}$, $j = 1, \cdots, p$ being selected, search for $\hat{h}_n$ that minimizes the bandwidth criteria in use. $h_n = \hat{h}_n$ is chosen to be the bandwidth that provides estimates of the norm of the varying coefficients.

5. Fix $h_{20} = 0.5\hat{h}_1$, $h_{2j} = \hat{h}_{2j}$, $j = 1, \cdots, p$ and $h_n = \hat{h}_n$. Search for $\hat{h}_l$ that minimizes the bandwidth criteria in use. $h_l = \hat{h}_l$ is selected to be the bandwidth for the estimation of the monotonic link function.

5.2 Nearest Neighbour Bandwidth

In this section, the Cross-Validation criterion is applied as a means of varying bandwidth selection. We intend to use the idea of nearest neighbour bandwidth. Suppose M is a positive integer. At a particular observation point, the first M adjacent observations are used as the local data for statistical exploration. M here determines the expand of the bandwidth

In terms of estimating the varying coefficients, the bandwidth selection takes place with respect to U. At any given point $U = u$, denote the distance between observation points U_i and u as D, and

$$D_i = ||U_i - u||, \quad i = 1, \cdots, n.$$

Sort the distance vector D in an increasing order gives

$$\{D_{(1)}, \ldots, D_{(n)}\}.$$

The observations are also sorted with respect to absolute distance. The sorted observations are

$$\{U_{(1)}, \mathbf{X}_{(1)}, y_{(1)}; \ldots ; U_{(n)}, \mathbf{X}_{(n)}, y_{(n)}\}$$

Then the first M nearest observations, $U_{(i)}, \mathbf{X}_{(i)}, y_{(i)}$, $i = 1, \cdots, M$, are used as effective information at point u. Denote by $r = D_{(M)}$, and the kernel function $K_r\left(D_{(i)}\right)$ at $U_{(i)}$ is defined as $K\left(\frac{D_{(i)}}{r}\right)/r$.

Similarly, when the target is to estimate the unknown link function, the varying bandwidth selection is based on $t = \mathbf{X}^T\hat{\boldsymbol{\beta}}(U)$. Denote M_l as a positive integer that determines the bandwidth span. At a given location t_0, denote the distance between observation t_i and t_0 by D, and

$$D_i = ||t_i - t_0||, \quad i = 1, \cdots, n$$

Sort the distances in an increasing order gives

$$\{D_{(1)}, \ldots, D_{(n)}\}.$$

Then sorted the observations in order with respect to distance. The sorted observations are

$$\{t_{(1)}, y_{(1)}; \ldots ; t_{(n)}, y_{(n)}\}.$$

Then the first M_l closest observations are used as effective information at point t_0. Denote by $r = D_{(M_l)}$, and the kernel function $K_r\left(D_{(i)}\right)$ at $U_{(i)}$ is defined as $K\left(\frac{D_{(i)}}{r}\right)/r$.

Cross-Validation Criterion

When one-step method is used for the estimation of the directions of the varying coefficients, the Cross-Validation objective function for model (3.1) is defined as

$$CV(M_1, M_n, M_l) = -2\sum_{i=1}^{n} l\left(\hat{g}^{-i}({\mathbf{x}_i}^T \hat{\boldsymbol{\beta}}_{-i}(U_i), y_i)\right). \qquad (5.5)$$

Similarly, when the directions of the varying coefficients are estimated with two-step method, the objective function is

$$CV(M_{21}, \cdots, M_{2p}, M_n, M_l) = -2\sum_{i=1}^{n} l\left(\hat{g}^{-i}({\mathbf{x}_i}^T \hat{\boldsymbol{\beta}}_{-i}(U_i), y_i)\right), \qquad (5.6)$$

where M_1,$M_{21}, \cdots, M_{2p}$, M_n and M_l are positive integers that determines the local neighbourhood for estimation of the directions, the unknown norm and the monotonic link function, respectively. $\hat{\boldsymbol{\beta}}_{-i}(\cdot)$ and $\hat{g}^{-i}(\cdot)$ are the estimators of $\boldsymbol{\beta}(U_i)$ and $g(\cdot)$, which are computed with the i_{th} observation been deleted. A combination of M_1,$M_{21}, \cdots, M_{2p}$, M_n and M_l that minimizes corresponding CV criterion is then searched.

An ideal bandwidth selection algorithm which is potential in balancing the computational cost and the bandwidth selection performance is desirable. As is introduced in the previous section, the proposed bandwidth selection algorithm should be working on this. When the nearest neighbour bandwidth is of interest, the thesis suggests to look for proper bandwidths using the same algorithm.

5.3 Simulations

In terms of real world application, data-driven bandwidths are more desirable, instead of depending on a researcher's experience. In our generalized varying coefficient models, there are quite a few bandwidths to be considered, which are the one-step estimation bandwidth h_1, the two-step estimation bandwidths h_{2j}, $j = 1, \cdots, p$, for different components of the varying coefficients, the bandwidth h_n for estimating the unknown norm and the bandwidth h_l for estimating the unknown link function. This means that the data-driven bandwidth selection could be very difficult. In addition to the difficulty of handling the data-driven bandwidths, there is another issue to be considered, that is the computation cost. The best choice is to not chase the practically optimal, which gives a combination of all different bandwidths that provides the best estimation performance, but the practically satisfactory, which trades off between accuracy and time cost, instead. To provide with a demonstration, 100 simulations with sample size $n = 200$ $n = 400$ and $n = 800$ for two examples are conducted. With both the Akaike information (AIC) and the Cross-Validation (CV) criteria proposed, the thesis has obtained data-driven global constant bandwidths and nearest neighbour varying bandwidths, respectively.

- **Example 1** $\beta_1(U) = sin(2\pi U)$; $\beta_2(U) = cos(2\pi U)$.
- **Example 2** $\beta_1(U) = sin(3\pi U)$; $\beta_2(U) = cos(2\pi U)$.

5.3.1 Data-driven constant bandwidth

The selection of data-driven constant bandwidth is provided by AIC and CV criteria. To balance the bandwidth selection and computation cost, the proposed algorithms are used to search for a combination of practical bandwidths. Table 12 demonstrates proper estimation results of the varying coefficients for both the AIC and the CV criteria. The AIC criterion provides slightly better estimation performance than the CV criterion, while the advantage is not dramatic.

Example 1	n=200		n=400		n=800	
AIC	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0216	0.0168	0.0118	0.007	0.0076	0.0039
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0216	0.016	0.0109	0.064	0.0058	0.003
CV	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0226	0.0173	0.013	0.008	0.0078	0.0039
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0245	0.0171	0.0122	0.0068	0.0061	0.0032

Example 2	n=200		n=400		n=800	
AIC	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0614	0.0618	0.0356	0.0345	0.0224	0.0227
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0634	0.0643	0.0399	0.0344	0.0226	0.0214
CV	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0658	0.0647	0.0359	0.0356	0.0227	0.0227
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0706	0.0663	0.0409	0.0366	0.0231	0.0221

Table 12: Estimation with data-driven constant bandwidths.

5.3.2 Data-driven varying nearest bandwidth

By inquiring CV criteria, the data-driven varying nearest bandwidths are also selected. Again both one-step and two-step estimation method to obtain the estimators of the directions of the varying coefficients are applied and compared. The bandwidth selection procedure is identical as that used for data-driven constant bandwidth in precious section. The thesis is not going to repeat here. Estimation results shown in Table 13 concludes with reasonable estimation performance of the varying coefficients for both examples, when CV criterion is used.

Example 1	n=200		n=400		n=800	
CV	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0318	0.0276	0.0176	0.0141	0.0098	0.0062
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0321	0.028	0.0187	0.0104	0.0082	0.0047

Example 2	n=200		n=400		n=800	
CV	1-step	2-step	1-step	2-step	1-step	2-step
$MISE\{\hat{\beta}_1(\cdot)\}$	0.0745	0.0739	0.0425	0.0401	0.0287	0.0276
$MISE\{\hat{\beta}_2(\cdot)\}$	0.0801	0.0793	0.0479	0.0427	0.0293	0.0289

Table 13: Estimation with data-driven nearest neighbour bandwidths.

6 Confidence band and hypothesis test

Confidence bands construction and hypothesis testing are two important subjects that statistical inferences are interested in. Construction of confidence bands is the major advantage of the bootstrap method (Faraway,1990). This thesis frequently utilizes the bootstrap approach for the construction of confidence bands and the hypothesis test.

6.1 Confidence bands of the varying coefficients

In practice, the intention is sometimes to estimate the confidence band of the coefficients. The construction of confidence band is based on the distribution of the maximum discrepancy between the estimated functional coefficient and the true functional coefficient. It is hard to find the exact distribution of the maximum discrepancy, however, it can be estimated by bootstrap.

For the j_{th} varying coefficient $\beta_j(\cdot)$, $j \in \{1, \cdots, p\}$, consider the quantity

$$T_j = \sup_{u \in U} \frac{|\hat{\beta}_j(u) - \beta_j(u)|}{\{var(\hat{\beta}_j(u)|\mathbf{X}, U)\}^{1/2}},$$

where $\hat{\beta}_j(\cdot)$ is the estimator of $\beta_j(\cdot)$, and $var(\hat{\beta}_j(\cdot)|\mathbf{X}, U)$ is the conditional variance of $\hat{\beta}_j(\cdot)$. Suppose the upper α quantile of T_j is c_α . If c_α and $\{var(\hat{\beta}_j(\cdot)|\mathbf{X}, U)\}^{1/2}$ in T_j were both known, the confidence band could be constructed as

$$\hat{\beta}_j(u) \pm \{var(\hat{\beta}_j(u)|\mathbf{X}, U)\}^{1/2} c_\alpha. \tag{6.1}$$

Now the problem is to estimate c_α and $\{var(\hat{\beta}_p(u)|\mathbf{X}, U)\}^{1/2}$, respectively. Denote the corresponding estimators by $\hat{c}_\alpha$ and $\{var^\star(\hat{\beta}_p(u)|\mathbf{X}, U)\}^{1/2}$, then the confidence band is constructed as

$$\hat{\beta}_j(u) \pm \{var^\star(\hat{\beta}_j(u)|\mathbf{X}, U)\}^{1/2}\hat{c}_\alpha. \tag{6.2}$$

The estimation of c_α and $\{var(\hat{\beta}_j(u)|\mathbf{X}, U)\}^{1/2}$ using the bootstrap approach is demonstrated as the following:

(1) With the proposed maximum rank correlation estimation method, estimate the functional coefficients $\boldsymbol{\beta}(\cdot)$. Denote the estimators for each functional coefficient as $\hat{\beta}_j(\cdot)$, $j \in \{1, \cdots, p\}$.

(2) For each $i = 1, \cdots, n$, generate a bootstrap sample $Y_i^\star$ based on the estimated log conditional density function

$$\ell\left[\hat{g}^{-1}\left\{\sum_{j=1}^{p} X_{ji}\hat{\beta}_j(U_i)\right\}, y\right].$$

Estimate the varying coefficients $\beta_j(\cdot)$ based on the bootstrap sample $\left\{\mathbf{X}^T, U, Y^\star\right\}$ using the same estimation method, *i.e.* the proposed maximum rank correlation estimation method. The resulting estimator $\hat{\beta}_j^\star(\cdot)$ is termed as a bootstrap sample of $\hat{\beta}_j(\cdot)$, $j \in \{1, \cdots, p\}$.

(3) Repeat step (2) M times provides M bootstrap samples $\hat{\beta}_{j,i}(\cdot)$, where $i = 1, \cdots, M$ and $j \in \{1, \cdots, p\}$. The estimator $\{var^\star(\hat{\beta}_j(\cdot)|\mathbf{X}^T, U)\}$ is taken to be the sample variance of the bootstrap sample $\hat{\beta}_{j,i}(\cdot)$, where $i = 1, \cdots, M$ and $j \in \{1, \cdots, j\}$.

(4) Repeat (2) another M_2 times to get bootstrap samples $\hat{\beta}_{j,i}(\cdot)$, where $i = 1, \cdots, M_2$ and $j \in \{1, \cdots, p\}$. For each varying coefficient of each bootstrap sample, compute the quantity

$$T_{j,i}^\star = \sup_{u \in U} \frac{|\hat{\beta}_j^\star(u) - \hat{\beta}_j(u|}{\{var^\star(\hat{\beta}_j(u)|X, U)\}^{1/2}},$$

where $i = 1, \cdots, M_2$ and $j \in \{1, \cdots, p\}$. $T^{\star}_{j,i}$, $i = 1, \cdots, M_2$ is termed as a bootstrap sample of T_j for the j_{th} coefficient function. The estimator $\hat{c}_\alpha$ is taken to be the upper α percentile of $T^{\star}_{j,i}$, $i = 1, \cdots, m$.

6.2 Hypothesis test of constant coefficients

The hypothesis test is another important part of statistical inference. The basic idea of most hypothesis test techniques is to compare the observed value of a test statistic with its empirical distribution calculated under the assumption that the null hypothesis were true. The null is then rejected or retained according to the magnitude of the test statistic relative to this distribution. However, as the test statistic is compared to an empirical distribution, rather than the true distribution that the test statistic follows. It is possible that the null hypothesis could be over-rejected or over-accepted.

Bootstrap hypothesis testing is a simulation-based testing method that involves re-sampling from the sample and the estimators and construction of simulated test statistic. In this study, of interest is to test whether a covariate impacts on the response variable constantly. The hypothesis is often constructed as

$$H_0 : \boldsymbol{\beta}_j(\cdot) = C_j \longleftrightarrow H_1 : \boldsymbol{\beta}_j(\cdot) \neq C_j, \tag{6.3}$$

where C_j, $j \in \{1, \cdots, p\}$ is the unknown constant coefficient indicating non-varying impact of the j_{th} covariate upon the response variable. The null hypothesis in (6.3) means the impact of the j_{th} covariate $\mathbf{X}_j$ is not varying over time, and the alternative hypothesis means the impact is time varying.

Without loss of generality, in this section, the hypothesis test are conducted using bootstrap approach for the p_{th} component of the coefficient vector. Consider the hypothesis

$$H_0 : \boldsymbol{\beta}_p(\cdot) = C_p \longleftrightarrow H_1 : \boldsymbol{\beta}_p(\cdot) \neq C_p. \tag{6.4}$$

The crucial quantity to be considered for the test is

$$T = \sup_{u \in U} \frac{|\hat{\beta}_p(u) - C_p|}{\{var(\hat{\beta}_p(u)|\mathbf{X}, U)\}^{1/2}},$$

where $\hat{\beta}_p(\cdot)$ is the estimator of the functional coefficient $\beta_p(\cdot)$ and $var(\hat{\beta}_p(\cdot)|\mathbf{X}, U)$ is the conditional variance of $\hat{\beta}_p(\cdot)$. The null hypothesis which attributes the p_{th} covariate with constant coefficient C_p

Suppose the upper α quantile of T under null hypothesis of (6.4) is c_α . If c_α, C_p and $\{var(\hat{\boldsymbol{\beta}}_p(u)|\mathbf{X}, U)\}^{1/2}$ in T were all known, the hypothesis test with the rejection region can be defined as

$$\sup_{u \in U} \frac{|\hat{\beta}_p(u) - C_p|}{\{var(\hat{\beta}_p(u)|\mathbf{X}, U)\}^{1/2}} > c_\alpha.$$

The problem is to estimate c_α, C_p and $\{var(\hat{\beta}_p(\cdot)|X, U)\}^{1/2}$, respectively. Denote by the corresponding estimators as $\hat{c}_\alpha$, $\hat{C}_p$ and $\{var^\star(\hat{\beta}_p(\cdot)|X, U)\}^{1/2}$, the hypothesis with size α is rejected when

$$\sup_{u \in U} \frac{|\hat{\beta}_p(u) - \hat{C}_p|}{\{var^\star(\hat{\beta}_p(u)|\mathbf{X}, U)\}^{1/2}} > \hat{c}_\alpha.$$

The null hypothesis is accepted otherwise. Now the thesis illustrates how to estimate c_α, C_p and $\{var(\hat{\beta}_p(\cdot)|\mathbf{X}, U)\}^{1/2}$.

C_p can be simply estimated by taking average of the estimates $\hat{\beta}_p(\cdot)$ across the datum points as

$$\hat{C}_p = n^{-1} \sum_{i=1}^{n} \hat{\beta}_p(U_i).$$

The thesis then demonstrates the estimation of c_α and $\{var(\hat{\beta}_p(\cdot)|\mathbf{X}, U)\}^{1/2}$ using the bootstrap approach.

(1) Under the null hypothesis, treat all other coefficients as functional and the p_{th} coefficient as constant, and estimate the functional coefficients and the constant coefficient, respectively. Denote the estimators by $\bar{\beta}_j(\cdot)$, $j \in \{1, \cdots, p-1\}$, and $\hat{C}_p$.

(2) For each $i = 1, \cdots, n$, generate a bootstrap sample $Y_i^\star$ based on the estimated log conditional density function

$$\ell\left[\hat{g}^{-1}\left\{\sum_{j=1}^{p-1} X_{ji}\bar{\boldsymbol{\beta}}_j(U_i) + X_{pi}\hat{C}_p\right\}, y\right].$$

Treat $\beta_p(\cdot)$ as functional and estimate it based on the bootstrap sample $\left\{\mathbf{X}^T, U, Y^\star\right\}$. The resulting estimator $\hat{\beta}_p^\star(\cdot)$ is termed as a bootstrap sample of $\hat{\beta}_p(\cdot)$

(3) Repeat (2) M times provides M bootstrap samples $\hat{\beta}_{p,i}(\cdot)$, $i = 1, \cdots, M$. The estimator $var^\star(\hat{\beta}_p(\cdot)|X, U)$ is taken to be the sample variance of the bootstrap sample $\hat{\beta}_{p,i}(\cdot)$, $i = 1, \cdots, M$.

(4) Repeat (2) another M_2 times to get a bootstrap sample $\hat{\boldsymbol{\beta}}_{p,i}(\cdot)$, $i = 1, \cdots, M_2$. And compute

$$T_i^\star = \sup_{u \in U} \frac{|\hat{\beta}_p^\star(u) - \hat{C}_p|}{\{var^\star(\hat{\beta}_p(u)|X, U)\}^{1/2}}.$$

$T_i^\star$, $i = 1, \cdots, M_2$ is termed as a bootstrap sample of T. The estimator $\hat{c}_\alpha$ is taken to be the upper α percentile of $T_i^\star$, $i = 1, \cdots, M_2$.

6.3 Bandwidth selection

Bandwidth selection is always an issue to be considered in this section. In the model settings, the varying coefficients are estimated in a few stages. In each stage, there are corresponding bandwidth selection problems to be solved. To look for the optimal practical bandwidths, complex algorithm and time consuming computation are two obstacles that are hard to demise. In previous simulation studies, the thesis has attempted to supply a bandwidth selection algorithm that is not too complex and not too computationally expensive. However, the observed sample and bootstrap samples are subject to different bandwidths when applicable. We do not want to repeat the above bandwidth selection algorithm for each and every bootstrap sample. Bootstrap approach itself is indeed computationally expensive. It would be exhausting if one attempts to search for data-driven bandwidths for each and every bootstrap sample. These bandwidths can not simply be given according to experience neither. To reduce the effort in bandwidth selection, the thesis attempts to use data-driven bandwidths for bootstrap samples with as less computational cost as possible.

Our intention of bandwidth selection is two-stage. First of all, for 1000 bootstrap samples, corresponding bandwidths are chosen by the AIC criterion. Then, bandwidths used for hypothesis test and confidence bands construction are determined to be the average of the corresponding bandwidths for these 1000 bootstrap samples. The reason here is that with these bandwidths, there would be a balance between computation saving and statistical validity. Details of these two-stage bandwidth selection are as follows.

Suppose the observed data set is proceeded with the proposed maximum rank correlation estimation method where the directions of the varying coefficients are estimated with the two-step method. The bandwidth used in the first stage of the two-step method is denoted by h_{20}. Since the second stage bandwidth is not sensitive to the first stage bandwidth, h_{20} is used for all bootstrap samples when the target is to estimate the directions of the varying coefficients. Denote by h_l the bandwidth for link function estimation when the observed data set is treated. Since the link function is not the main target for the hypothesis test and confidence band construction, h_l is used for bandwidth selection for bootstrap samples.

1.) For bootstrap sample $\{Y_i^\star, \mathbf{X}_i, U_i\}$, $i = 1, \cdots, n$. with h_{20} and $h_{2j} = h_{20}$ for $j = 1, \cdots, p$, estimates of the directions of the varying coefficients are obtained. Search for $\tilde{h}_n$ that minimizes the AIC criterion.

2.) With h_{20} and $\tilde{h}_n$, search for $h_{2j}, j = 1, \cdots, p$, that minimizes the AIC criterion. Denote these bandwidths as $\hat{h}_{2j}, j = 1, \cdots, p$.

3.) $\tilde{h}_n$ is not the practical optimal bandwidth for the estimation of the norm. With fixed h_{20} and $\hat{h}_{2j}, j = 1, \cdots, p$, search for the bandwidth for estimating the norm again, and the resulting bandwidth is denoted by $hath_n$.

4.) Repeat the above steps 100 times gives 100 sets of bandwidths. To make the presentation more clear, express these bandwidths as $h_{2j,i}, j = 1, \cdots, p$, and $h_{n,i}$ for $i = 1, \cdots, 100$. The combination of bandwidths used for upcoming bootstrap samples are taken to be

$$h_{2j} = \frac{1}{100} \sum_{i=1}^{100} h_{2j,i}, \quad j = 1, \cdots, p,$$

and

$$h_n = \frac{1}{100} \sum_{i=1}^{100} h_{n,i}.$$

6.4 Simulation: Hypothesis test

In this section, the thesis implements the hypothesis test within the framework of Poisson regression. The link function is selected to be the commonly used *log* transformation function for data generation.

- **Example** $\beta_1(u) = sin(\pi u) + 0.8; \qquad \beta_2(u) = cos(\pi u) - 0.4.$

For this example with support $u \in [0, 1]$, the null and alternative hypotheses are

$$H_0 : \beta_j(\cdot) = C_j \longleftrightarrow H_1 : \beta_j(\cdot) \neq C_j,$$

where $j = 1, 2$. They test whether the impact of the covariates on the response variable are varying or constant.

To test the impact of the first covariate, set

$$\beta_1(u) = bsin(\pi u) + 0.8 \text{ and } \beta_2(u) = cos(\pi u) - 0.4,$$

where $b \in \{0, 0.1, \cdots, 1\}$. For each fixed b, 1000 simulations with sample size $n = 500$ are conducted. For significance level $\alpha = 0.05$ and $\alpha = 0.1$, the hypothesis

$$H_0 : \beta_1(\cdot) = C_1 \longleftrightarrow H_1 : \beta_1(\cdot) \neq C_1$$

is tested and the power of the test for each b is evaluated.

The impact of the second covariate is tested in the similar way. Set

$$\beta_1(u) = sin(\pi u) + 0.8 \text{ and } \beta_2(u) = (1 - b) + bcos(\pi u) - 0.4,$$

where $b \in \{0, 0.1, \cdots, 1\}$. For each fixed b, 1000 simulations with sample size $n = 500$ are conducted. For significance level $\alpha = 0.05$ and $\alpha = 0.1$, the hypothesis

$$H_0 : \beta_2(\cdot) = C_2 \longleftrightarrow H_1 : \beta_2(\cdot) \neq C_2$$

is tested, and the power functions are calculated.

To examine how powerful the hypothesis tests are, the estimated power functions for the varying coefficients are shown in Figure 9, which suggests proper performance of the bootstrap approach.

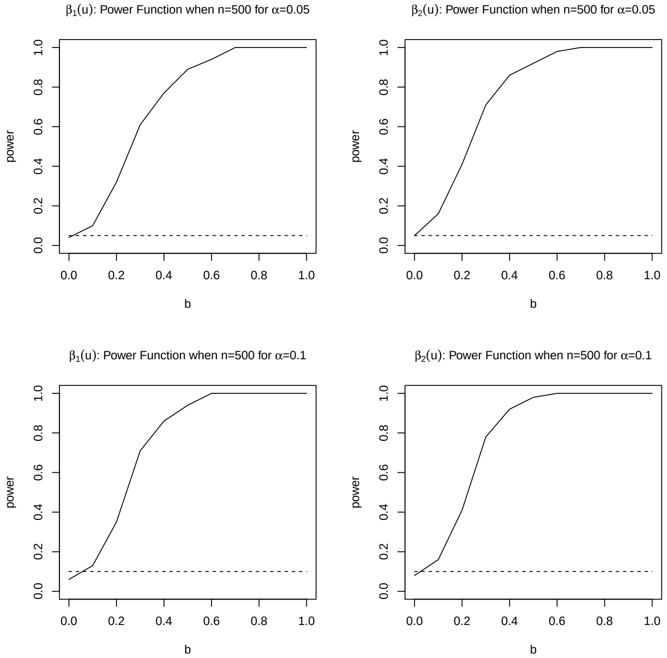


Figure 9: Power function.
The figures on the left are power functions of the hypothesis for $\beta_1(\cdot)$ for significance level $\alpha = 0.05$ and 0.1, respectively;The figures on the right are power functions of the hypothesis for $\beta_2(\cdot)$ for significance level $\alpha = 0.05$ and 0.1, respectively.The dotted lines highlight the corresponding significance level.

6.5 Simulation: Confidence bands construction

Suppose it has been identified that all coefficients are functional. For the following two examples, confidence bands of the varying coefficients are constructed using the bootstrap approach.

- **Example 1** $\beta_1(u) = sin(\pi u);$ $\beta_2(u) = cos(2\pi u).$

Coverage probability	$\beta_1(\cdot)$	$\beta_2(\cdot)$
1-α=99 %	90%	94%
1-α=95 %	87%	92%

Table 14: Example 1: Coverage probability.

For example 1 with support $u \in [0, 1]$. We construct the bootstrap confidence bands with the confidence level $1 - \alpha$ is taken to be 99% and 95%, respectively. The bandwidths in the estimation procedure, provided by AIC criterion, are $h_{20} = 0.0622, h_{21} = 0.1789, h_{22} = 0.0979, h_n = 0.1448, h_l = 0.1208$, respectively. When it comes to the bootstrap confidence bands construction for each simulation, the corresponding bandwidths for bootstrap samples are derived from 100 bootstrap samples using the same AIC-based bandwidth selection algorithm. The Monte Carol errors are of size $\sqrt{0.95 * 0.05/500} \approx 0.01$ for $\alpha = 0.05$ and $\sqrt{0.99 * 0.01/500} \approx 0.004$ for $\alpha = 0.05$. Table 14 shows that there are poor coverage probabilities.

- **Example 2** $\beta_1(u) = sin(2\pi u);$ $\beta_2(u) = cos(2\pi u).$

Coverage probability	$\beta_1(\cdot)$	$\beta_2(\cdot)$
1-α=99 %	88%	95%
1-α=95 %	85%	93%

Table 15: Example 2: Coverage probability.

Fo example 2 with support $u \in [0, 1]$, the bootstrap confidence bands with confidence level $1 - \alpha$ 99% and 95% are constructed. The bandwidths in the estimation procedure ,provided by AIC criterion, are $h_20 = 0.0531, h_{21} = 0.2511, h_{22} = 0.2573, h_n = 0.3259, h_l = 0.1136$; respectively. In the process of bootstrap confidence bands construction for each simulation, the corresponding bandwidths for bootstrap samples are derived from 100 bootstrap samples using the same AIC-based bandwidth selection algorithm. In Table 15, the coverage probabilities confirm that current attempt of confidence bands construction is potentially non-satisfactory.

Discussion and an explortory approach

Through exploration of the variance of the directions and the norm. It is found worthy of constructing confidence bands for the varying coefficients through a composition of corresponding confidence bands for the directions and the norm.

Estimation of the directions involves an standardization operation, due to that the norm of the direction is fixed to be 1. Such operation has the following tendency: when the absolute value of an estimate $\hat{\beta}_j(\cdot)$ at u is large, say $|\hat{\beta}_j(u)| \to 1$, the variance of the estimate would tend to 0. Therefore, in the attempt of achieving $var^\star(\hat{\beta}_{0p}(u))$, the bootstrap samples $\hat{\beta}^\star_{0p}(u)$ of $\hat{\beta}_{0p}(u)$ are estimated with the two-step estimation method without standardization operation. Standardized $\hat{\beta}^\star_{0p}(\cdot)$ are used in all other occasions when relevant.

Denote $[B_0^{low}(\cdot),B_0^{up}(\cdot)]$ and $[N^{low}(\cdot),N^{up}(\cdot)]$ the confidence bands for the directions and the norm, where the superscripts *low* and *up* indicate lower and upper confidence bands, respectively. Let

$$B(\cdot) = \left\{B_0^{low}(\cdot)N^{low}(\cdot), B_0^{up}(\cdot)N^{low}(\cdot), B_0^{low}(\cdot)N^{up}(\cdot), B_0^{up}(\cdot)N^{up}(\cdot)\right\}.$$

The confidence bands of the varying coefficients at any given point u is constructed as

$$[min\,(B(u))\,, max\,(B(u))].$$

With this sort of confidence bands composition, example 1 and 2 with sample size $n = 500$ are treated, respectively. Bandwidths selection is proceeded as instructed previously.

- **Example 1** $\beta_1(u) = sin(\pi u)$; $\beta_2(u) = cos(2\pi u)$.

Coverage probability	$\beta_1(\cdot)$	$\beta_2(\cdot)$
1-α=99 %	97%	98%
1-α=95 %	95%	96%

Table 16: Example 1: Improved coverage probability

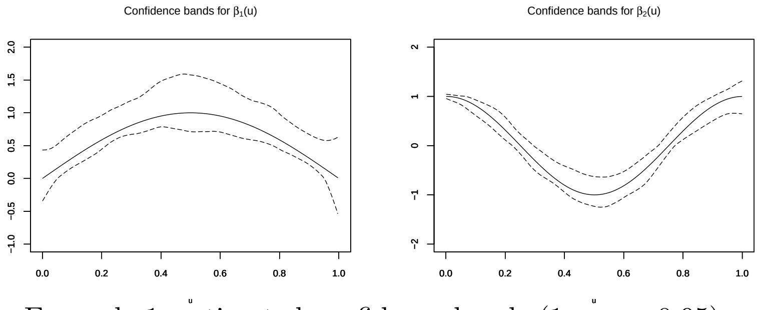


Figure 10: Example 1: estimated confidence bands $(1 - \alpha = 0.95)$:

The dotted lines are the confidence bands with medium average band width, and the solid lines are the true curves.

Table 16 indicates that the coverage probabilities have been dramatically improved to a satisfactory level. Figure 10 presents estimated confidence bands for $\beta_1(\cdot)$ and $\beta_2(\cdot)$, which have the medium band width. Compare with previously constructed confidence bands, the idea of composition gives more smooth and reliable confidence bands.

- **Example 2** $\beta_1(u) = sin(2\pi u)$; $\beta_2(u) = cos(2\pi u)$.

Coverage probability	$\beta_1(\cdot)$	$\beta_2(\cdot)$
1-α=99 %	97%	98%
1-α=95 %	96%	96%

Table 17: Example 2: Improved coverage probability

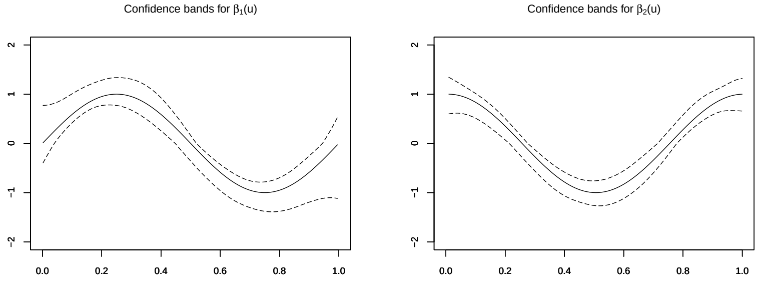


Figure 11: Example 2: estimated confidence bands $(1 - \alpha = 0.95)$.

The dotted lines are the confidence bands with medium average band width, and the solid lines are the true curves.

The Monte Carol errors are of size 0.01 for $\alpha = 0.01$ and 0.004 for $\alpha = 0.05$. The coverage probabilities have too been significantly improved as well. Figure 11 depicts estimated confidence bands for $\beta_1(\cdot)$ and $\beta_2(\cdot)$, which have the medium width of the confidence bands. Through constructing confidence bands with the composition of confidence bands for the directions and the norm, reliable confidence bands are obtained again.

7 Extention to Generalized Semi-Varying Coefficient Models with Unknown Monotonic Link Transformation

The thesis has explored generalized varying coefficient models in previous sections. In this section, the thesis extends the research to generalized semi-varying coefficient models. In practice, some coefficients in generalized varying coefficient models may be constant. We pay a price on the variance side of an estimator of a constant component, when the constant component is treated as functional. The estimation for a generalized semi-varying coefficient model is straightforward. However, it prompts the question of how to identify the composition of the coefficient vector, among which there are both varying and constant components. This is basically a model selection problem.

7.1 Generalized Semi-varying Coefficient Models

Suppose that, a generalized semi-varying-coefficient model, whose model structure, $\mathcal{M}_L(p,q)$, involves L covariates among which p of them have varying compact upon the response variable, and $q = L - p$ of them have constant effect on the response variable. The mean regression function is supposed to be linear via a monotonic link function $g(\cdot)$ as

$$g\{m(U,\mathbf{X},\mathbf{Z})\} = \mathbf{X}^T\boldsymbol{\beta}(U) + \mathbf{Z}^T\boldsymbol{\alpha}, \tag{7.1}$$

where $\boldsymbol{\beta}(\cdot) = \{\beta_1(\cdot),\ldots,\beta_p(\cdot)\}$ is the varying-coefficient vector, p-dimensional $\mathbf{X}$ and q-dimentional $\mathbf{Z}$ are the covariates, and $\boldsymbol{\alpha} = \{\alpha_1,\ldots,\alpha_q\}$ is the vector holding constant coefficients. Without loss of generality, the thesis assumes that the link function $g(\cdot)$ is a strictly increasing link transformation.

For the purpose of further presentation, denote the coefficient vector holding both varying and constant components by

$$\mathcal{B}(\cdot) = \{\boldsymbol{\beta}(\cdot)^T, \boldsymbol{\alpha}^T\}^T,$$

and the norm of the coefficient vector by $N(\cdot) = \|\mathcal{B}(\cdot)\|$. Further, the direction of $\mathcal{B}(\cdot)$ is denoted by

$$\mathcal{B}_0(\cdot) = \{\boldsymbol{\beta}_0(\cdot)^T, \boldsymbol{\alpha}_0(\cdot)^T\}^T,$$

where $\mathcal{B}_0(\cdot) = \frac{\mathcal{B}(\cdot)}{\|\mathcal{B}(\cdot)\|}$, $\boldsymbol{\beta}_0(\cdot) = \frac{\boldsymbol{\beta}(\cdot)}{\|\boldsymbol{\beta}(\cdot)\|}$, and $\boldsymbol{\alpha}_0(\cdot) = \frac{\boldsymbol{\alpha}(\cdot)}{\|\boldsymbol{\alpha}(\cdot)\|}$. To deal with the identifiability issue, set $\|\mathcal{B}(0)\| \equiv 1$. Let $\mathbf{H}^T = \left\{\mathbf{X}^T, \mathbf{Z}^T\right\}^T$, (7.1) is equivalent to

$$g\{m(U, \mathbf{H})\} = \mathbf{H}^T \mathcal{B}(U). \tag{7.2}$$

7.2 Estimation procedure

Before proposing strategic model identification, the thesis introduces the estimation procedures first, assuming that the true model structure is successfully identified.

7.2.1 Estimation of the directions of the coefficients

First of all, estimators of the directions of the coefficients have to be achieved. The estimation of the directions of the coefficients can be free from the model structure, because the direction of a constant coefficient is also varying as long as the norm is not constant. Since the estimation of the direction has nothing to do with the model structure, there is no difference between a generalized varying coefficient model and a generalized semi-varying coefficient model, in terms of direction approximation. We can estimate the coefficient vector $\mathcal{B}_0(\cdot)$ with either a one-step or a two-step method proposed previously. Since their is nothing new in the estimation for the directions of the coefficients, detailed presentation of the estimation procedure is simply recalling of previous proposed method, which we are not going to repeat here.

7.2.2 Estimation of the varying coefficients

When estimates of the directions of the coefficients are obtained, the norm of the coefficient vector is estimated using the method introduced to generalized varying coefficient models. Denote the norm of the coefficient vector $\|\mathcal{B}(\cdot)\|$ by $N(\cdot)$. Let

$$v = \mathbf{H}^{\mathrm{T}}\hat{\mathcal{B}}_0(U), \text{ and } v_i = \mathbf{h}^{\mathrm{T}}\hat{\mathcal{B}}_0(U_i),$$

where $\hat{\mathcal{B}}_0(U) = \{\hat{\boldsymbol{\beta}}_0^T(U), \hat{\boldsymbol{\alpha}}_0^T(U)\}^T$ is the estimator of the directions of the coefficients by either one-step or two-step estimation method. Replacing the directions of the varying coefficients by their estimators gives

$$\mathbf{h}_i^{\mathrm{T}}\mathcal{B}_0(U_i)N(U_i) \approx \mathbf{h}_i^{\mathrm{T}}\hat{\mathcal{B}}_0(U_i)N(U_i),$$

which yields the following rank correlation between y and $vN(U)$

$$\sum_{i \neq j} I(y_i > y_j) I\left(z_i N(U_i) > z_j N(U_j)\right).$$

For any given u, given U_i is in a small neighbourhood of u, by the Taylor's expansion

$$N(U_i) \approx N(u) + \dot{N}(u)(U_i - u),$$

the local rank correlation can be approximated by

$$\sum_{i \neq j} I(y_i > y_j) I\left(z_i \left\{N(u) + \dot{N}(u)(U_i - u)\right\} > z_j \left\{N(u) + \dot{N}(u)(U_j - u)\right\}\right) \times K_{h_n}(U_i - u) K_{h_n}(U_j - u),$$

where $K_{h_n}(t) = K(t/h_n)/h_n$, $K(t)$ is the kernel function, and h_n is the smoothing parameter defining the width of the neighbouring at $U = u$.

Because $N(u) > 0$, the above objective function is equivalent to

$$\sum_{i \neq j} I(y_i > y_j) I\left(z_i \{1 + c(u)(U_i - u)\} > z_j \{1 + c(u)(U_j - u)\}\right) \times K_{h_n}(U_i - u) K_{h_n}(U_j - u), \qquad (7.3)$$

where $c(u)$ corresponds to $\dot{N}(u)/N(u)$. Let $\hat{c}(u)$ maximise (7.3), $\hat{c}(u)$ is an estimator of $\dot{N}(u)/N(u)$, and the estimator of $N(u)$ is generated by

$$\hat{N}(u) = \exp\left\{\int_0^u \hat{c}(u) du\right\}. \qquad (7.4)$$

The estimator of $\boldsymbol{\beta}(u)$ is therefore conducted via

$$\hat{\boldsymbol{\beta}}(u) = \hat{N}(u)\hat{\boldsymbol{\beta}}_0(u).$$

7.2.3 Estimation of the constant coefficients

When the estimators of the direction of the constant coefficient vector and the norm of the coefficient vector are in hand. The estimator for a constant coefficient is straightforward. Firstly, treat the constant coefficient vector as varying, *i.e.* $\boldsymbol{\alpha}(U) = \boldsymbol{\alpha}$. Denote the estimator of this varying version as

$$\tilde{\boldsymbol{\alpha}}(U_i) = \hat{\boldsymbol{\alpha}}_0(U_i)\hat{N}(U_i).$$

The estimator of the constant coefficient vector is achieved by simplify taking average of the estimates of $\boldsymbol{\alpha}(\cdot)$ across the observations of the scaler from U_1 to U_n, *i.e.*

$$\hat{\boldsymbol{\alpha}} = n^{-1}\sum_{i=1}^{n}\tilde{\boldsymbol{\alpha}}(U_i).$$

7.3 Identification of constant coefficients

With the estimation procedure being proposed, the question is how to identify which coefficients are constant and which are variable. In this thesis, cross-validation (CV) and Akaike information criterion (AIC) for model selection are applied to identify the constant coefficients. The model selection algorithms are provided by Zhang(2011).

CV based model identification

For each $i = 1, \cdots, n$, delete the ith observation and estimate the coefficients using the method proposed. Denote the estimators as $\hat{\boldsymbol{\beta}}^{-i}(\cdot)$ and $\hat{\boldsymbol{\alpha}}^{-i}$, then the log conditional density function of y at i given$\mathbf{X}_i$, $\mathbf{Z}_i$ and U_i is

$$L_i = logf(y_i; \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Z}_i^T\hat{\boldsymbol{\alpha}}^{-i}, \mathbf{X}_i^T\hat{\boldsymbol{\beta}}^{-i}(U_i))$$

Thus the cross validation score is given by

$$CV = -n^{-1} \sum_{i=1}^{n} L_i$$

To compute the CV scores for all possible models is not practically feasible. For a model with in total L covariates, there are 2^L possible models to be considered. Of interest is to reduce the computational burden on one hand, and retain the model selection accuracy on the other. Suppose the underline true model structure is $M_{p,q}$, $p + q = L$,where p and q determine the number of varying coefficients and constant coefficients. This section demonstrates the Backward elimination and the Discrepancy from average algorithms proposed by Zhang (2013).

• **Backward elimination**

Instead of computing the CV scores for all possible models, the backward elimination operates as follows:

1. Start by playing a full model $M_{L,0}$, with all coefficients are set to be varying. The corresponding CV is computed and denoted by CV_L

2. Treat one of the covariates $\beta_j(\cdot)$, $j = 1, \cdot, L$ to be constant, with others being varying. This involves in total L models. Compute their CV scores as CV_{L-1_j}, $j = 1, \cdot, L$. Denote the smallest CV score by $CV_{L-1} = CV_{L-1_k}$. Suppose this score is achieved by model $M_{L-1,1}$ with the kth coefficient being constant. This model is considered a candidate model.

3. If $CV_{L-1} > CV_L$, the final model is selected to be $M_{L,0}$. Otherwise, with the kth coefficient being fixed as constant, repeat step 2 by treating $L - 1$ models with one more constant coefficient. Decision of model structure is made by comparing the smallest CV score CV_{L-2} with CV_{L-1}. If $CV_{L-2} > CV_1$, then the model is selected to be $M_{L-1,1}$. Otherwise, continue

the backward elimination.

A backward elimination algorithm would treat in maximum $2^{-1}L(L+1)$ CV computations, which reduces the computation largely from the identification of all possible models.

• **Discrepancy from average**

A more time saving way for the model selection is based on the discrepancy from average algorithm, which focus on the discrepancy of the estimated function from its average. The algorithm operates as follows:

1. Again, start by playing a full model $M_{L,0}$. The corresponding CV is computed and denoted by CV_L. Then compute the discrepancy of the estimated function $\hat{\beta}_j\cdot$, $j = 1, \cdots, L$ from its average as

$$d_j = \sum_{i=1}^{n} \{\hat{\beta}_j(U_i) - \bar{\beta}_j\}^2, \text{ with } \bar{\beta}_j = n^{-1} \sum_{i=1}^{n} \hat{\beta}_j(U_i).$$

Sort d_j, $j = 1, \cdots, L$, in an increasing order as $d_{(1)} \le d_{(2)} \le \cdots \le d_{(L)}$. Suppose the possibility that a coefficient function is constant is in the same order as $d_{(j)}$, $j = 1, \cdots, L$

2. Treat the (1)-st coefficient,which has discrepancy $d_{(1)}$, as constant, with others being varying. Model $M_{L-1,1}$ is viewed as a candidate model, and its corresponding CV is computed as CV_{L-1}. If $CV_{L-1} > CV_L$, then the model selection is ended, and the identified model structure is $M_{L,0}$. Otherwise, continue the model identification by treating the (2)-nd coefficient as constant.

The discrepancy from average algorithm is less accurate than the backward elimination algorithm, however, the number of models it is to investigate is only to the maximum L.

AIC based model identification

For each $i = 1, \cdots, n$, estimate the coefficients using the method proposed. Denote the estimators as $\hat{\boldsymbol{\beta}}(\cdot)$ and $\hat{\boldsymbol{\alpha}}$, then the log conditional density function of y given $\mathbf{X}_i$, $\mathbf{Z}_i$ and U_i is

$$L_i = logf(y_i; \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Z}_i^T\hat{\boldsymbol{\alpha}}, \mathbf{X}_i^T\hat{\boldsymbol{\beta}}(U_i)).$$

Thus the AIC score can be calculated via

$$AIC = -2\sum_{i=1}^{n} L_i + 2K,$$

where K is the number of parameters involved in our estimation procedure.

$$K = q + 0.6h_1^{-1}p + 1.285714(h_n^{-1} + h_l^{-1}),$$

when the directions are estimated with one-step method; and

$$K = q + \sum_{j=1}^{p} 0.6h_{2j}^{-1} + 1.285714(h_n^{-1} + h_l^{-1}),$$

when the directions are estimated with two-step method.

• **Backward elimination**

1. Start by treating the model as a generalized varying coefficient model,$M_{L,0}$, with all coefficients are varying. In such case, the corresponding AIC score is

$$AIC_L = -2\sum_{i=1}^{n} logf(Y_i; X_i, Z_i, Z_i^T\hat{\boldsymbol{\alpha}}(U_i), X_i^T\hat{\boldsymbol{\beta}}(U_i)) + 2K,$$

where

$$K = 0.6h_1^{-1}L + 1.285714(h_n^{-1} + h_l^{-1}),$$

when the directions are estimated with one-step method; and

$$K = \sum_{j=1}^{L} 0.6 h_{2j}^{-1} + 1.285714(h_n^{-1} + h_l^{-1}),$$

when the directions are estimated with two-step method.

2. Reduce the number of varying coefficients by one, and denote the version of model as $M_{L-1,1}$. For each possible model of this version, the AIC is constructed by treating one of the varying coefficients as constant. The model with minimal AIC score as is selected as a candidate model with AIC_{L-1}. If $AIC_{L-1} > AIC_L$, stop the model identification, and model $M_{L,0}$ is selected. Otherwise, continue the model identification.

3. Suppose now the model is $M_{L-k,k}$ with AIC_{L-k} (If $k = L$, the model idetification procedure is terminated, and the model $M_{0,L}$, with all coefficients are constant, is selected.). Reduce the number of varying coefficients by one, and denote the version of model as $M_{L-k-1,k+1}$. For each possible model of this version, the AIC is constructed by treating one of the varying coefficients as constant. The model with minimal AIC score as is selected as a candidate model with AIC_{L-k-1}. If $AIC_{L-k-1} > AIC_{L-k}$, stop the model identification, and model $M_{L-k,k}$ is selected. Otherwise, continue the model identification.

• **Discrepancy from average** For the discrepancy from average algorithm, the identification procedure is identical to the backward elimination algorithm, except that in each step, only one candidate model is considered.

1. Start with the full model $M_{L,0}$. Calculate the AIC score and denote it by AIC_L.

2. Suppose at present the model is $M_{L-k,k}$ with AIC_{L-k} (If $k = L$, the model idetification procedure is terminated, and the model $M_{0,L}$, with

all coefficients are constant, is selected.). Among the varying coefficients in this model, let the one with minimal discrepancy from average as constant gives a candidate model $M_{L-k-1,k+1}$ with AIC_{L-k-1}. If $AIC_{L-k-1} > AIC_{L-k}$, terminate the model identification, and model $M_{L-k,k}$ is selected. Otherwise, continue the model identification.

7.4 Simulation studies

Simulation studies in this section implement the model selection and estimation for generalized semi-varying coefficient models with respect to two examples.

7.4.1 Model Identification

Model selection is different from estimation. Therefore, the thesis is heading for a different approach to bandwidths. In Zhang's (2013) work, as long as the bandwidths used are not ridiculously small, it is suggested to use bandwidths as small as possible. However, the situation becomes different in this project.

There are quite a few bandwidths to be handled simultaneously, and those bandwidths are inter-dependent. Even the obtaining of the log-likelihood function requires an estimation of the link function. Previous simulation studies have demonstrated that our method does give proper estimators for the directions of the varying coefficients, regardless of the model structure. Therefore, it is confident to use small bandwidth for the estimation of the directions, in terms of model selection. However, the estimation of the unknown norm depends heavily on the estimator of its derivative.

When the identification of model structure is of interest, it is not necessary to spend too much effort in computation. Therefore, only the one-step estimation method with small bandwidths $h_1 = 0.05$ is used for the estimation of directions of the coefficients, and bandwidth $h_n = 0.05, 0.1, 0.15$ and 0.2 for estimating the norm are used for evaluation.

Example 1 $\beta_1(u) = sin(2\pi u)$; $\beta_2(u) = cos(\pi u) - 0.2$; $\alpha_1 = 0.$; $\alpha_2 = 0.6$.

Example 2 $\beta_1(u) = sin(2\pi u)$; $\beta_2(u) = cos(2\pi u) - 1 + \sqrt{0.8}$; $\alpha_1 = -0.2$; $\alpha_2 = 0.4$.

For example 1 and example 2, CV and AIC criteria with both the Backward elimination and Discrepancy from average algorithms are applied for model selection. The frequencies that the right models are picked up using the backward elimination algorithm are recorded in Table 18

Algorithm	Eg. 1		Eg. 2	
Backward elimination	CV	AIC	CV	AIC
$hn = 0.05$	82	64	96	98
$hn = 0.1$	90	91	98	99
$hn = 0.15$	90	92	99	100
$hn = 0.2$	90	92	99	100

Table 18: The number of picking up the right model among 100 attempts.

It is seen that although it is suggested to use small bandwidth for model structure identification, one has to be careful in the decision of how small a bandwidth should be. With different bandwidths being used for the estimation of the norm, it is noticed that the frequency that the true model structure is identified correctly tends to be stable as the bandwidths increases. Therefore, this thesis suggests to try out a few different bandwidths for the norm and believe in the model structure that most of the attempts agree.

With the bandwidth for estimating the norm being fixed as $h_n = 0.15$, the thesis identifies the underline model structure for example 1 and example 2 with both the Backward elimination and the discrepancy from average algorithm, the frequencies that the right model are identified are recorded in Table 19

$h_n = 0.15$	Eg. 1		Eg. 2	
Algorithm	CV	AIC	CV	AIC
Backward elimination	90	92	99	100
Discrepancy from average	82	83	91	93

Table 19: The number of picking up the right model among 100 attempts.

It is seen that one has more than 90% chance of picking up the true model structure when using the backward elimination algorithm, and more than 85% with discrepancy from average algorithm. This suggests that the proposed model selection methods work well.

7.4.2 Estimation for GSVCMs

The thesis is going to estimate for generalized semi-varying coefficient functions with respect to two examples. Firstly, treat all coefficients as varying and the directions of the coefficients are estimated. Both the one-step and the two-step estimation methods are applied and their estimation performances are compared. The estimator of the constant coefficients are derived by taking average of its varying version over $U_i, i = 1, \cdots, n$. At this stage, the bandwidths used are the practical optimal bandwidths that minimizes corresponding mean integrated squared errors.

• **Example 1** $\beta_1(u) = sin(2\pi u)$; $\beta_2(u) = cos(\pi u) - 0.2$; $\alpha_1 = 0.$; $\alpha_2 = 0.6$.

	n=200		n=400		n=800	
	1-step	2-step	1-step	2-step	1-step	2-step
MISE($\hat{\beta}_1(\cdot)$)	0.024	0.021	0.015	0.013	0.009	0.008
MISE($\hat{\beta}_2(\cdot)$)	0.021	0.015	0.011	0.009	0.008	0.006
MISE($\hat{\boldsymbol{\alpha}}_1(\cdot)$)	0.004	0.004	0.002	0.002	0.001	0.001
MISE($\hat{\boldsymbol{\alpha}}_2(\cdot)$)	0.01	0.01	0.006	0.006	0.003	0.003

Table 20: Example 1: Mean integrated squared errors.

For Example 1, 100 simulations with sample size $n = 200$, $n = 400$ and $n = 800$ are conducted. Table 20 demonstrates the reasonable estimation performance for both the varying and constant coefficients. As sample size increases, the estimation accuracy improves steadily. For the varying coefficients, the two-step estimation method surpasses the one-step method in estimation accuracy. Whereas, for the constant coefficients, it is noticed that the one-step and the two-step estimation methods are comparable.

- **Example 2** $\beta_1(u) = sin(2\pi u)$; $\beta_2(u) = cos(2\pi u) - 1 + \sqrt{0.8}$; $\alpha_1 = -0.2$; $\alpha_2 = 0.4$.

	n=200		n=400		n=800	
	1-step	2-step	1-step	2-step	1-step	2-step
MISE($\hat{\beta}_1(\cdot)$)	0.024	0.018	0.015	0.01	0.009	0.006
MISE($\hat{\beta}_2(\cdot)$)	0.022	0.017	0.013	0.009	0.006	0.004
MISE($\hat{\boldsymbol{\alpha}}_1(\cdot)$)	0.006	0.006	0.003	0.003	0.002	0.002
MISE($\hat{\boldsymbol{\alpha}}_2(\cdot)$)	0.005	0.005	0.003	0.003	0.002	0.002

Table 21: Example 2: Mean integrated squared errors.

For Example 2, 100 simulations with sample size $n = 200$, $n = 400$ and $n = 800$ are conducted as well. Table 21 indicates that the estimation performance for both the varying and constant coefficients are reasonable.

As sample size increases, the estimation error decreases steadily. It is also found that the two-step estimation method performs better than the one-step method in estimation accuracy for the varying coefficients. For the constant coefficients, the one-step and the two-step methods give comparable estimation results.

7.4.3 Estimation with data-driven bandwidths

At this stage, the thesis is to estimate for generalized semi-varying coefficient function with data-driven bandwidths. Since the two-step estimation method is practically more applicable than the one-step method, only the two-step estimation method is used for the estimation of the directions of the varying coefficients.

For Example 1 and Example 2, 100 simulations with sample size $n = 200$, $n = 400$ and $n = 800$ are conducted, respectively. Data-driven bandwidth with both the AIC and CV algorithms are used and compared. Estimation results are shown in Table 22, which demonstrates that the proposed estimation method work reasonably well. With the two-step estimation method being applied, the AIC and CV criteria produce comparable estimation results.

Example 1	n=200		n=400		n=800	
Criterion	AIC	CV	AIC	CV	AIC	CV
MISE($\hat{\beta}_1(\cdot)$)	0.027	0.027	0.013	0.013	0.008	0.008
MISE($\hat{\beta}_2(\cdot)$)	0.028	0.029	0.014	0.013	0.008	0.007
MISE($\hat{\boldsymbol{\alpha}}_1(\cdot)$)	0.004	0.004	0.002	0.002	0.001	0.001
MISE($\hat{\boldsymbol{\alpha}}_2(\cdot)$)	0.014	0.013	0.007	0.006	0.004	0.004

Example 2	n=200		n=400		n=800	
Criterion	AIC	CV	AIC	CV	AIC	CV
MISE($\hat{\beta}_1(\cdot)$)	0.025	0.025	0.014	0.014	0.008	0.008
MISE($\hat{\beta}_2(\cdot)$)	0.029	0.029	0.015	0.015	0.007	0.007
MISE($\hat{\boldsymbol{\alpha}}_1(\cdot)$)	0.007	0.007	0.004	0.004	0.003	0.003
MISE($\hat{\boldsymbol{\alpha}}_2(\cdot)$)	0.008	0.008	0.007	0.007	0.006	0.005

Table 22: Estimation with data-driven bandwidths.

8 Extension to panel data

The main interest of this research lies in studying the association between the covariates and the response variable under the setting of generalized varying coefficient models with unknown monotonic link function. In this section, the thesis extends the exploration to panel data. In many applications, data from different subjects are collected over a period of time. The number of data points and the time of data collection for each subject might be different. Suppose the number of subjects involved in a study is N. Let $(\mathbf{X}_s^{\mathrm{T}},\ U_s,\ Y_s)$ be an i.i.d. sample collected for the s_{th} subject, $s = 1, \cdots, N$. $\mathbf{X}_s^T$ is p-dimensional covariates, and U_s is a scalar that the variation of the impact of the covariates depend on. For simplicity, only univariate U_s is considered in this thesis.

For the s_{th} subject, denote by $\mathbf{X}_{si}^T$, U_{si} and Y_{si}, $i = 1, \cdots, n_s$, the set of observations of the corresponding covariates, scalar and responsible variable, where n_s, $s = 1, \cdots, N$, is the number of observations collected for the s_{th} subject. Then the generalized varying coefficient model with unknown monotonic link function for the s_{th} subject is constructed as

$$g\left\{m(\mathbf{X}_{s,i}, U_{s,i})\right\} = \mathbf{X}_{s,i}^{\mathrm{T}}\boldsymbol{\beta}_s(U_{s,i}), \tag{8.1}$$

and

$$\boldsymbol{\beta}_s(U_{si}) = \boldsymbol{\beta}(U_{s,i}) + \boldsymbol{\epsilon}_s, \text{ with } \boldsymbol{\epsilon} \in N(\mathbf{0}_p, \sigma\mathbf{I}_{p\times p}), \tag{8.2}$$

where $m(\mathbf{X}_{s,i}, U_{s,i}) = E(Y_{s,i}|\mathbf{X}_{s,i}, U_{s,i})$ is the mean regression function, $\boldsymbol{\beta}_s(\cdot)$ is the p-dimensional unknown functional vector holding the varying coefficient functions, $\boldsymbol{\beta}(\cdot)$ is the p-dimensional unknown functional vector holding the common trend and $\boldsymbol{\epsilon}_s$ is a p-dimensional error term that captures individual difference. Without loss of generality, assume that $\|\boldsymbol{\beta}(0)\| = 1$. We endeavour to apply the method proposed in previous sections for the estimation of $\boldsymbol{\beta}(\cdot)$ and the unknown monotonic link function $g(\cdot)$.

8.1 Estimation Procedure

In fact model (8.1) is not identifiable, since the norm of $\boldsymbol{\beta}_s(U_{si})$ is unknown and different for each subject. However, one does have access to $\boldsymbol{\beta}_s(U_{si})$ in any way. The true vector varying coefficient function $\boldsymbol{\beta}_s(U_{si})$ defined in (8.2) is equivalent to

$$\mathbf{B}_s(U_{s,i})||\boldsymbol{\beta}_s(0)|| = \boldsymbol{\beta}(U_{s,i}) + \boldsymbol{\epsilon}_s,$$

where $\mathbf{B}_s(\cdot) = \frac{\boldsymbol{\beta}_s(\cdot)}{||\boldsymbol{\beta}_s(0)||}$ can be estimated using the proposed maximum rank correlation estimation method given that $||\mathbf{B}_s(0)|| = 1$. However,since $||\boldsymbol{\beta}_s(0)||$ is unknown, one has no access to the estimation of $\boldsymbol{\beta}(\cdot)$ at the moment. It is realized that, it is possible find a path to get an insight of $\boldsymbol{\beta}(\cdot)$ via setting control subject.

Without loss of generality, treat the first subject to be a control subject and derive the following

$$\frac{\mathbf{B}_s(U_{si})||\boldsymbol{\beta}_s(0)||}{||\boldsymbol{\beta}_1(0)||} = \frac{\boldsymbol{\beta}(U_{si})}{||\boldsymbol{\beta}_1(0)||} + \frac{\boldsymbol{\epsilon}_s}{||\boldsymbol{\beta}_1(0)||}.$$

Denote $K_s = \frac{||\boldsymbol{\beta}_s(0)||}{||\boldsymbol{\beta}_1(0)||}$, and $K = \frac{1}{||\boldsymbol{\beta}_1(0)||}$, then

$$\mathbf{B}_s(U_{si})K_s = \boldsymbol{\beta}(U_{si})K + \boldsymbol{\epsilon}_s K,$$

where both K_s and K are positive definite. Denote further $\boldsymbol{\alpha}(\cdot) = \boldsymbol{\beta}(\cdot)K$, and $\boldsymbol{\mu}_s = \boldsymbol{\epsilon}_s K$, which lead to

$$\mathbf{B}_s(U_{si})K_s = \boldsymbol{\alpha}(U_{si}) + \boldsymbol{\mu}_s.$$

The above transformations makes the estimation of $\boldsymbol{\alpha}(\cdot)$, which is related to $\boldsymbol{\beta}(\cdot)$ with an unknown constant factor K, possible, since $\boldsymbol{\mu} \in N(\mathbf{0}_p, K\sigma\mathbf{I}_{p\times p})$. Suppose the estimates $\hat{\mathbf{B}}_s(\cdot)$ and $\hat{K}_s$ are obtained, the estimator of $\boldsymbol{\alpha}(\cdot)$ is taken to be

$$\hat{\boldsymbol{\alpha}}(\cdot) = \frac{1}{N}\sum_{s=1}^{N} \hat{\mathbf{B}}_s(\cdot)\hat{K}_s.$$

The problem now lies in the estimation of the K_s. We propose an approach as the following. Since all subjects are from the same group which share a common $\boldsymbol{\beta}(\cdot)$, the true varying coefficient functions are parallel, *i.e.*

$$\boldsymbol{\beta}_s(\cdot) - \boldsymbol{\beta}_1(\cdot) = \boldsymbol{\epsilon}_s - \boldsymbol{\epsilon}_1,$$

which is equivalent to

$$\mathbf{B}_s(\cdot)K_s - \mathbf{B}_1(\cdot)K_1 = \mathbf{B}_s(\cdot)K_s - \mathbf{B}_1(\cdot) = \boldsymbol{\mu}_s - \boldsymbol{\mu}_1.$$

Suppose now estimates of $\hat{\mathbf{B}}_s(\cdot)$ are obtained. Denote the distance between $\hat{\mathbf{B}}_{s,d}(\cdot)K_s$ and $\hat{\mathbf{B}}_{1,d}(\cdot)$ by

$$\mathbf{C}_d(\cdot) = \hat{\mathbf{B}}_{s,d}(\cdot)K_s - \hat{\mathbf{B}}_{1,d}(\cdot),$$

the mean value of the distance $\mathbf{C}_d(\cdot)$ by

$$\mathbf{e}_d = \frac{1}{n_1}\sum_{i=1}^{n_1} \mathbf{C}_d(U_{1,i}),$$

and the variation of $\mathbf{C}_d(\cdot)$ from the mean value $\mathbf{e}_d$ by

$$\mathbf{v}_d = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\mathbf{C}_d(U_{1,i}) - \mathbf{e}_d)^2.$$

Then the objective function (8.3)

$$L(K_s) = \sum_{d=1}^{p} \mathbf{v}_d \tag{8.3}$$

is considered, and the estimator of K_s is chosen to be the minimizer of (8.3).

8.2 Simulation studies

The number of subjects is set to be $N = 20$. For each subject Y_s, $s = 1, \cdots, N$, the observations are independently collected at $U_{s,i}$, for $i = 1, \cdots, n_s$, where n_s is the number of observations of the s_{th} subject. The proposed maximum rank correlation estimation method can not afford difficulties caused by a small number of sample size. To insure that, at each panel, the varying coefficients can be properly estimated, n_s, $s = 1, \cdot, N$, are drawn from Uniform distribution $Unif[0.8n, n]$, where $n = 500$ is the designed number of observations. U has uniform support $[0, 1]$. The covariates $\mathbf{X}_{s,i}$, for $i = 1, \cdots, n_s$, are two-dimensional and independently drawn from the standard normal distribution, and the error vector $\boldsymbol{\epsilon}_s$, is two-dimensional and independently drawn from the normal distribution $N(\mathbf{0}_2, 0.5\mathbf{I}_{p\times p})$.

8.2.1 Estimation for the panel data

- **Example 1:** $\beta_{s,1}(u) = sin(\pi u) + \epsilon_{s,1}$; $\beta_{s,2}(u) = cos(2\pi u) + \epsilon_{s,2}$; $s = 1, \cdots, N$.

For Example 1, 100 simulations with maximum number of observations $n = 200, 400$ and 800 are conducted. Recall the definitions:

$$\boldsymbol{\beta}_s(U_{si}) = \mathbf{B}_s(U_{si})||\boldsymbol{\beta}_s(0)|| = \boldsymbol{\beta}(U_{si}) + \boldsymbol{\epsilon}_s,$$

and

$$\mathbf{B}_s(U_{si})K_s = \boldsymbol{\alpha}(U_{si}) + \boldsymbol{\mu}_s,$$

where $K_s = \frac{\boldsymbol{\beta}_s(0)||}{||\boldsymbol{\beta}_1(0)||}$, and $K = \frac{1}{||\boldsymbol{\beta}_1(0)||}$. $\mathbf{B}_s(\cdot)$, K_s and $\boldsymbol{\alpha}(\cdot)$ are estimated, respectively. Bandwidths are selected by considering MISE as functional to corresponding bandwidth. The MISEs for the estimation of individual varying coefficients are calculated by assuming that $||\boldsymbol{\beta}_s(0)||$ is known. $||\boldsymbol{\beta}_s(0)||$ is absorbed by the link function which will not be involved in the estimation procedure.

MISE of	$\{\hat{\beta}_{s,01}(\cdot)\}$	$\{\hat{\beta}_{s,02}(\cdot)\}$	$\{\hat{\beta}_{s,1}(\cdot)\}$	$\{\hat{\beta}_{s,2}(\cdot)\}$
s= 1	0.0023	0.0054	0.0155	0.0244
s= 2	0.0093	0.006	0.011	0.0197
s= 3	0.0054	0.005	0.0102	0.034
s= 4	0.0073	0.0056	0.016	0.0229
s= 5	0.003	0.0052	0.0145	0.0199
Group	$MISE\{\hat{\beta}_1(\cdot)\} =$	0.0041	$MISE\{\hat{\beta}_2(\cdot)\} =$	0.0096

Table 23: Example 1: Mean integrated squared errors ($n = 800$).

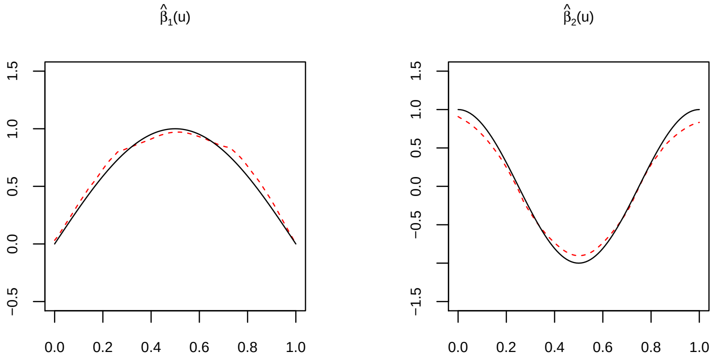


Figure 12: Example 1: Estimated varying coefficients ($n = 800$).
The solid lines are the true functional curves.The dotted lines are the estimated curves.

For simplicity, the thesis only present MISE outcomes for the first 5 subjects with sample size $n = 800$. The mean integrated squared errors in table 23 indicate proper estimation of the varying coefficient at each panel. At this stage of the paper, all subjects are assumed to share the common varying coefficients $\boldsymbol{\beta}(\cdot)$. The estimation of $\boldsymbol{\beta}(\cdot)$ given $||\boldsymbol{\beta}_s(0)||$ is also satisfactory. In Figure 12, the estimated varying coefficients with medium estimation error among 100 simulations are depicted to give an insight into the estimation performance.

- **Example 2:** $\beta_{s,1}(u) = sin(2\pi u) + \epsilon_{s,1}$; $\beta_{s,2}(u) = cos(2\pi u) + \epsilon_{s,2}$; $s = 1, \cdots, N$.

100 simulations with maximum number of observations $n = 200, 400$ and 800 are conducted for example 2. The simulations estimate $\mathbf{B}_s(\cdot)$, K_s and $\boldsymbol{\alpha}(\cdot)$, respectively. Bandwidths are selected by considering MISE as functional to corresponding bandwidth. The MISEs for the estimation of individual varying coefficients are calculated by assuming that $||\boldsymbol{\beta}_s(0)||$ is known.

MISE of	$\{\hat{\beta}_{s,01}(\cdot)\}$	$\{\hat{\beta}_{s,02}(\cdot)\}$	$\{\hat{\beta}_{s,1}(\cdot)\}$	$\{\hat{\beta}_{s,2}(\cdot)\}$
s= 1	0.0043	0.0031	0.0301	0.0256
s= 2	0.0042	0.0046	0.0391	0.0318
s= 3	0.0035	0.0026	0.014	0.0139
s= 4	0.0041	0.003	0.0235	0.0191
s= 5	0.0045	0.0026	0.0204	0.0199
Group	$MISE\{\hat{\beta}_1(\cdot)\} =$	0.0136	$MISE\{\hat{\beta}_2(\cdot)\} =$	0.0143

Table 24: Example 2: Mean integrated squared errors ($n = 800$).

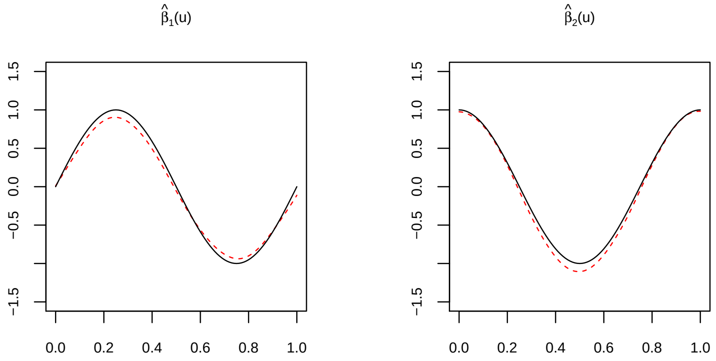


Figure 13: Example 2: Estimated varying coefficients ($n = 800$).
The solid lines are the true functional curves.The dotted lines are the estimated curves.

The mean integrated squared errors with sample size $n = 800$ in table 24 as well demonstrates satisfactory estimation of the varying coefficient at each panel. The estimated varying coefficients with medium estimation error among 100 simulations are ploted in Figure 13 as a demonstration of the estimation performance.

8.2.2 Estimation with data-driven bandwidths

At this stage, the thesis attempts to let the bandwidths be selected by the data itself. For simplicity, only the AIC criterion is applied. For Example 1 and Example 2, 100 simulations with sample size $n = 200$, $n = 400$ and $n = 800$ are conducted. Only estimation results for the first four subjects with sample size $n = 800$ are presented.

Example 1 $\beta_{s,1}(u) = sin(\pi u) + \epsilon_{s,1}$; $\beta_{s,2}(u) = cos(2\pi u) + \epsilon_{s,2}$; $s = 1, \cdots, 20$.

Panel	s = 1	s = 2	s = 3	s = 4
$MISE\{\hat{\beta}_{s,1}(\cdot)\}$	0.0173	0.012	0.011	0.0181
$MISE\{\hat{\beta}_{s,2}(\cdot)\}$	0.0275	0.0259	0.0376	0.0249
Group	$MISE\{\hat{\beta}_1(\cdot)\}$ =	0.006	$MISE\{\hat{\beta}_2(\cdot)\}$ =	0.0128

Table 25: Example 1: Estimation with AIC bandwidths ($n = 800$).

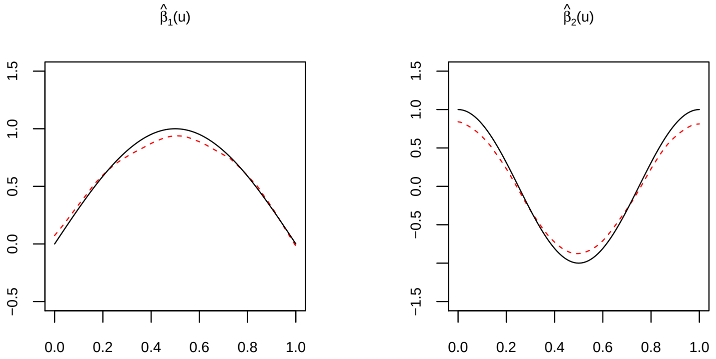


Figure 14: Example 1: Data-driven varying coefficients ($n = 800$).
The solid lines are the true functional curves.The dotted lines are the estimated curves.

Example 2 $\beta_{s,1}(u) = sin(2\pi u) + \epsilon_{s,1}$; $\beta_{s,2}(u) = cos(2\pi u) + \epsilon_{s,2}$; $s = 1, \cdots, 20$.

Panel	s = 1	s = 2	s = 3	s = 4
$MISE\{\hat{\beta}_{s,1}(\cdot)\}$	0.0373	0.0555	0.0155	0.0276
$MISE\{\hat{\beta}_{s,2}(\cdot)\}$	0.033	0.0458	0.0152	0.0214
Group	$MISE\{\hat{\beta}_1(\cdot)\} \quad = \quad 0.0188$		$MISE\{\hat{\beta}_2(\cdot)\} \quad = \quad 0.0184$	

Table 26: Example 2: Estimation with AIC bandwidths ($n = 800$).

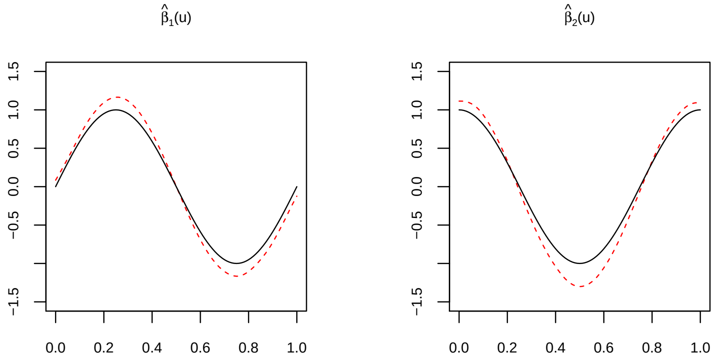


Figure 15: Example 2: Data-driven varying coefficients ($n = 800$).
The solid lines are the true functional curves.The dotted lines are the estimated curves.

The mean integrated squared errors with sample size $n = 800$ in Table 25 and Table 26 demonstrate reasonable estimation results of the varying coefficient both for each panel and the group. The estimation performance is depicted in Figure 14 and Figure 15, which plot the estimated varying coefficients with medium estimation error among 100 simulations.

9 Real Data Analysis: The Role of Visual Cues in The Mating Decisions of Satellite Horseshoe Crabs

At the new and full moon high tides during spring and summer, female horseshoe crabs together with their attached male partners arrive on the beach for fertilization (Brockmann and Penn, 1992; Brockmann, 1996). These paired crabs nest in the sand, where the males fertilize the eggs as the females lay them. Horseshoe crabs are highly male-biased (Brockmann, 1996). Apart from those males who are attached to females, unattached males come to the beach as well and crowd around the couples for chances of fertilization. These unattached male crabs are called satellites.

An interesting question is that whether the mating decisions of these unattached males are tactical or random. If satellite crabs randomly crowd around the paired couples, sizes of the groups should be similar (Brockmann, 1996) and females of pairs are likely to have equal values in fertilization (Brockmann, 1996; Schwab, 2006). However, observed group sizes are variable significantly, and attached females that attract satellites in general have greater fertilization values and lay more eggs than females that do not attract satellites. Since the mating decisions of satellite crabs are unlikely to be random, they are tactical.

There is evidence that male horseshoe crabs use visual cues to locate and pair with females or become their satellites (Barlow *et al.*, 1982,1987; Powers *et al.*, 1991; Brockmann, 1996; Passaglia *et al.*, 1997; Schwab and Brockmann, 2007), while ignoring others. Although more recent studies have identified that male crabs use other sensory cues, like chemical and tactile cues, to find their mates (Saunders *et al.*, 2010; Johnson and Brockmann, 2012) as well, it is still interesting to understand the role of visual cues in the mating decisions of satellite horseshoe crabs.

This thesis attempts to study the association between the number of satellites and visual cues (female age and size) of paired female horseshoe crabs using the proposed MRCE method. The problem refers to data from a study that investigated factors that affect whether the female crab had any satellites residing near her. The study was conducted by Brockmann (1996) and his colleagues and involved observations for 173 female horseshoe crabs. Each female horseshoe crab in this study had a male crab attached to her in her nest.

Denote Sa as the number of satellites residing around an attached female, and Sa_i: $i = 1, \cdots, 173$, as the number of satellites residing around the i^{th} female. Explanatory variables that are thought to affect satellite crabs' mating decisions include paired female crabs age and size. As female crabs grow older, their shell colour become darker, and spine conditions get worse. The field researchers categorized female crabs according to their colours and spine conditions, which are associate with female crabs' age. Paired female crabs' size was measured according to their weight (Wt) in kg and carapace width (W) in cm. Denote paired female crabs' colour by $C : 1 = \text{light medium}, 2 = \text{medium}, 3\text{=medium dark}, 4\text{=dark}$ and spine condition by $S : 1 = \text{both good}, 2 = \text{one worn or broken}, 3\text{=both worn or broken}$. The data is presented in Figure 16 and 17.

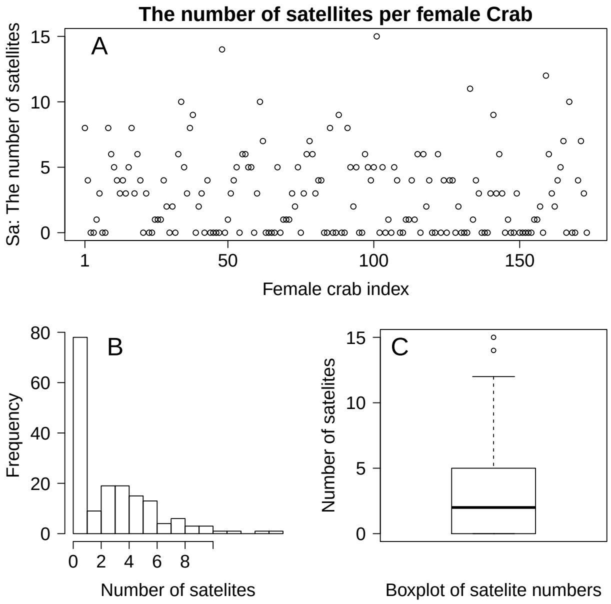


Figure 16: The number of satellites residing around paired female crabs.
(**A**): Sa - Observed numbers of satellites; (**B**): Histogram of Sa; (**C**): Boxplot of Sa.

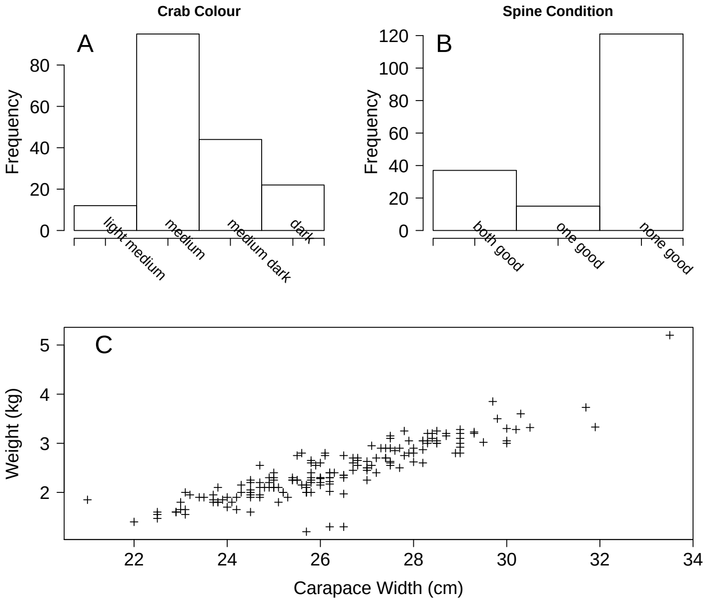


Figure 17: Size and Age of paired female crabs.
(**A**): C - Histogram of colour; (**B**): Histogram of spine condition; (**C**): Boxplot of Sa.

About 35% of the observed female crabs did not attract any satellite. However, these female crabs did not show less fertilization values, since they did not lay fewer eggs than female crabs that attracted satellites (Brockmann,1996). The unobserved difference in fertilization values between paired crabs that attracted satellites and those that did not might be a localized phenomenon in the observatory site. The paired female crabs were mostly mature adults that were neither too old nor too young. Around 80% of them were medium or medium dark, and more than 60% of them had both their spines worn or broken. Their spine conditions, in a way, demonstrate their success in surviving natural threats in the sea, since crabs that are too young or too old are more likely to be eliminated in natural selection. There is a clear linear association between weight and carapace width. Larger female crabs were heavier and more likely to have wider carapace width.

Colour(C) and spine condition (S) are categorical variables. Before proceeding to data analysis, they are transformed into dummy variables. Denote $\mathbf{C}^T = (C_1, C_2, C_3)$, where

$$\begin{aligned} C_1 &= 1, \quad \text{if light medium}, \quad C_1 = 0, \quad \text{otherwise;} \\ C_2 &= 1, \quad \text{if medium}, \quad C_2 = 0, \quad \text{otherwise;} \\ C_3 &= 1, \quad \text{if medium dark}, \quad C_2 = 0, \quad \text{otherwise,} \end{aligned}$$

and $\mathbf{S} = (S_1, S_2)$, where

$$\begin{aligned} S_1 &= 1, \quad \text{if both good}, \quad S_1 = 0, \quad \text{otherwise;} \\ S_2 &= 1, \quad \text{if one worn or broken}, \quad S_2 = 0, \quad \text{otherwise.} \end{aligned}$$

Let $\mathbf{X}^T = (\mathbf{C}^T, \mathbf{S}^T, W, Wt)$ be the covariate vector. The conditional mean regression function $m(\mathbf{X})$ is assumed to be linear via

$$g_t\left\{m(\mathbf{X})\right\} = \alpha + \mathbf{X}^T\boldsymbol{\beta}, \tag{9.1}$$

where $g_t(\cdot)$ is the unknown monotonic link function, α is the intercept term, and $\boldsymbol{\beta}$ is the vector holding coefficient parameters.

9.1 Maximum likelihood estimation

When count data is considered, the monotonic link function is frequently assumed to be the *log* function. When *log* transformation is applied, model (9.1) is specified as

$$log\left\{m(\mathbf{X})\right\} = a + \mathbf{X}^T\boldsymbol{\beta}. \tag{9.2}$$

With a standard MLE, the estimation results are presented in Table 27.

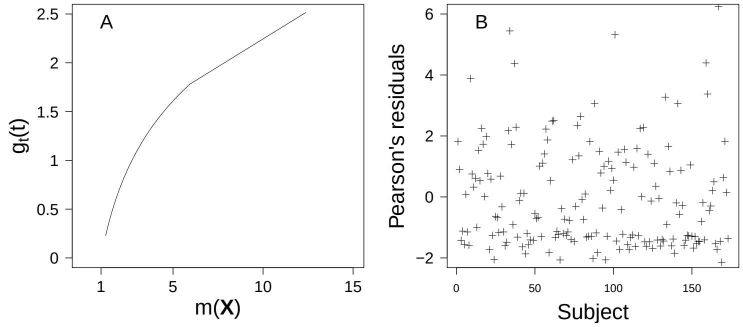
Table 27: Analysis with maximum likelihood estimation method.

AIC=-141.208 P-Residuals=533.779				
Parameter	Estimate	Std Error	95% CI	99% CI
a	-0.801	0.919	[-2.602, 1.000]	[-3.168, 1.566]
β_1	0.531	0.229	[0.082, 0.980]	[-0.059, 1.121]
β_2	0.266	0.162	[-0.052, 0.584]	[-0.151, 0.683]
β_3	0.018	0.183	[-0.341, 0.377]	[-0.453, 0.489]
β_4	-0.087	0.121	[-0.324, 0.150]	[-0.399, 0.245]
β_5	-0.238	0.202	[-0.634, 0.158]	[-0.758, 0.282]
β_6	0.017	0.048	[-0.077, 0.111]	[-0.107, 0.141]
β_7	0.497	0.166	[0.172, 0.822]	[0.069, 0.925]

The estimated coefficients indicate that, in the study, satellite male crabs tended to attach to those paired female crabs, which were lighter ($\hat{\beta}_1 = 0.531$, $\hat{\beta}_2 = 0.266$ and $\hat{\beta}_3 = 0.018$) and heavier ($\hat{\beta}_7 = 0.497$). Positive estimate

$\hat{\beta}_6 = 0.017$ reveals only limited role of carapace width in relation to the number of satellites. The association between satellite numbers and paired female crabs' spine condition was even negative ($\hat{\beta}_4 = -0.087, \hat{\beta}_5 = -0.238$). 95% Confidence intervals of the estimates emphasize the attractiveness of lighter colour and heavier weight for the satellites. Whereas, 99% confidence intervals only identify positive impact of female crabs' weight on the number of satellites.

Figure 18: MLE: Link function and residuals estimation.



(**A**): Estimates of the link function. (**B**): the Pearson's residuals.

It is interesting to see these estimates, since they might uncover how satellite crabs evaluate the age and size of paired female crabs. Compare with its spine conditions, the colour of a paired female might be a more credible indicator of age. Younger crabs are lighter in colour. They are more likely to (but not necessarily) have better spine conditions due to unforeseen lives under the sea surface, where young and old crabs face equal risks of survival. The size of female crabs is an important factor that determines fertilization values. Paired female crabs that attract satellites are larger and potentially more productive. However, MLE gives more explanatory power to weight than carapace width in relation to female crabs' size. The satellite crabs might evaluate the size of paired females in a more sophisticated way, other than relying too much on the carapace width.

9.2 Maximum rank correlation estimation

Due to that the intercept can be absorbed by the link function. Given $\mathbf{X}^T = (\mathbf{C}^T, \mathbf{S}^T, W, Wt)$, model (9.1) is equivalent to

$$g\left\{m(\mathbf{X})\right\} = \mathbf{X}^T \boldsymbol{\beta}. \tag{9.3}$$

To make the model identifiable, divide by $||\boldsymbol{\beta}||$ on both sides of model (9.3) and obtain

$$g_0\left\{m(\mathbf{X})\right\} = \mathbf{X}^T \boldsymbol{\beta}_0, \tag{9.4}$$

where $||\boldsymbol{\beta}_0|| = 1$ is the coefficient vector to be estimated. Thus $g_0(\cdot)$, relative to the true unknown link function $g_t(\cdot)$ is

$$g_0(\cdot) = \frac{g_t(\cdot) - \alpha}{||\boldsymbol{\beta}||}.$$

The transformation of the link function from $g_t(\cdot)$ to $g_0(\cdot)$ does not deteriorate monotonicity. Hence MRCE is applicable. Estimation results provided by MRCE are shown in Table 28.

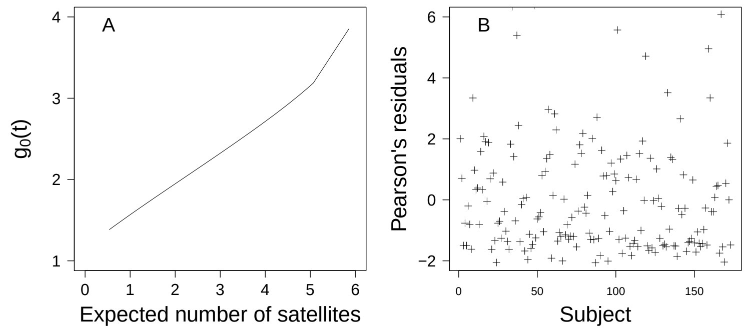
Table 28: Analysis with maxmum rank correlation estimation method.

		AIC=-147.485	P-Residuals=541.014	
Parameter	Estimate	Std Error	95% CI	99% CI
β_{01}	0.732	0.168	[0.403, 1.062]	[0.299, 1.165]
β_{02}	0.436	0.098	[0.243, 0.628]	[0.183, 0.689]
β_{03}	0.251	0.123	[0.009, 0.493]	[-0.067, 0.569]
β_{04}	0.014	0.136	[-0.252, 0.281]	[-0.336, 0.365]
β_{05}	-0.103	0.153	[-0.404, 0.197]	[-0.498, 0.291]
β_{06}	0.032	0.051	[-0.067, 0.132]	[-0.099, 0.163]
β_{07}	0.446	0.176	[0.100, 0.792]	[-0.009, 0.901]

According to biological definitions, colour and spine condition are two indicators of female crabs' age in their lifespan. Positive estimates of coefficients ($\hat{\beta}_{01} = 0.732$ and $\hat{\beta}_{02} = 0.436$ and $\hat{\beta}_{03} = 0.251$) suggest that female crabs, whose colour were lighter, were more attractive to male satellite crabs. Compare with colour, the spine condition was not a significant factor to satellites mating decision. In general, the MRCE supports that younger crabs were more admirable in the study field. The MLE method does not conclude with strong positive impact of colour in determining the satellite number, whereas the proposed MRCE method is more practically realistic than the MLE method. The confidence intervals provided by MRCE confirm the attractiveness of lighter colour of female crabs.

Estimates of coefficients for carapace width and weight are both positive, ($\hat{\beta}_{06} = 0.032$ and $\hat{\beta}_{07} = 0.446$). Bootstrap confidence intervals suggest positive association between female crabs' weight and satellite numbers. Compare with females' weights, their carapace widths possessed less fertilization values in the site of study. Up to this stage, MRCE is only confident in that unattached male crabs tend to join female crabs with lighter colour and heavier weight.

Figure 19: MRCE: Link function and residuals estimation (full model).



(**A**): Estimates of the link function. (**B**): the Pearson's residuals.

Compare with the MLE method, the MRCE method has smaller AIC score (-147.485<-141.208), and larger sum of squares of Pearson's residuals (541.014>533.779). Both methods find that larger female crabs are more attractive to male crabs. The proposed MRCE method further identifies positive fertilization values of female crabs' who were lighter in colour. The MRCE returns the estimated link, which is not the commonly used *log* function or its linear transformations. As the MRCE method is more practically realistic, this thesis would suggest to use MRCE method to fit the data, instead of the MLE method.

10 Real Data Analysis: Short-term Effects of Fluctuating Air Pollution on Health

Since the 1990s a plenty of studies have focused on the effects of air pollution on health outcomes in major European countries (Zmirou *et al.*, 1996; Sunyer *et al.*, 1996; Touloumi *et al.*, 1996; Wordley *et al.*, 1997), North America (Dockery *et al.*, 1993), South America (Saldiva *et al.*, 1995; Borja-Aburto *et al.*, 1997) and certain Asian cities (Xu *et al.*, 1994; Wong *et al.*, 1999). Positive associations between fluctuations of certain air pollutants and changes in mortalities or morbidities have been found in global cities, such as London (de Leon *et al.*, 1996), Amsterdam (Schouten *et al.*, 1996), Paris (Dab *et al.*, 1996), Milan (Vigotti *et al.*, 1996) and Hong Kong (Wong *et al.*, 1999; Wong *et al.*, 2002).

In these studies, the response variables are mostly health outcomes which are usually counts of hospital admissions for specific diseases. The core model is therefore frequently selected from Poisson regression models, such as *linear model* (Katsouyanni *et al.*, 1996,1997) or *generalized additive model* (Schwartz *et al.*, 1996; Schwartz, 1996;) with log transformation. There are at least two limitations in these studies: although the application of logarithmically transformed data is practically reasonable, it is not sure whether logarithmically transformed describes the data set correctly; the association between air pollution and health outcomes may vary over time, hence the scope should be extended from linear models to more complex models. In this thesis, we have introduced the MRCE method. The objective in this section is to extend from previous studies and implement the proposed method with data collected in Hong Kong.

10.1 Data description

The environmental data set consists of a collection of daily measurements of air pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995. The health outcomes are counts of daily hospital admissions of patients suffering from circulatory and respiratory diseases in the city. Surrounded by the South China Sea on the east, south, and west, Hong Kong is located on China's south coast. It has a humid subtropical climate with hot, humid summers and mild, dry winters. As an intensely urbanised and densely populated global city, local people in Hong Kong have experienced adverse effects of ambient air pollution (Wong *et al.*, 1999). Air pollutants that are considered of major importance include Sulphur Dioxide (SO_2), Nitrogen Dioxide (NO_2), Respirable suspended particulates (PM_{10}) and Ozone (O_3).

Denote date of measurements by t, daily average concentration levels of major air pollutants by covariates $\mathbf{X} = \{X_1, X_2, X_3, X_4\}$, where X_1 is SO_2 (in $\mu g/m^3$), X_2 is NO_2 (in $\mu g/m^3$), X_3 is PM_{10} (in $\mu g/m^3$) and X_4 is O_3 (in $\mu g/m^3$), and hospital daily circulatory admissions, daily respiratory admissions and daily total numbers of the two admissions as y_C, y_R and y, respectively. This section attempts to explore the association between changing air pollution levels and respiratory and circulatory health problems using the proposed MRCE method. We start by brief presentation of the data set.

10.1.1 Daily hospital admissions

According to International Classification of Diseases (ICD), target diseases were allocated into two groups: respiratory diseases (ICD 460-466, 471-478, 480-487, and 490-496) and circulatory diseases (ICD 410-417, 420-438, and 440-444) (Wong, *et al.*, 1999). Emergency hospital admissions for these two types of diseases were collected in all 12 major hospitals for 1994 and 1995. Although the study period was not very long, the numbers of daily hospital admissions were relatively high in 1994 and 1995 in Hong Kong. Summary statistics of health outcomes are depicted in Table 29 and Figure 20.

	Min	25th Percentile	Medium	Mean	75th Percentile	Max
Circulatory admissions	56	90	105	106.2	122	193
Respiratory admissions	87	128	150	153.2	174	285
Total admissions	152	222	253	259.4	291	450

Table 29: Summary statistics of daily hospital admissions.

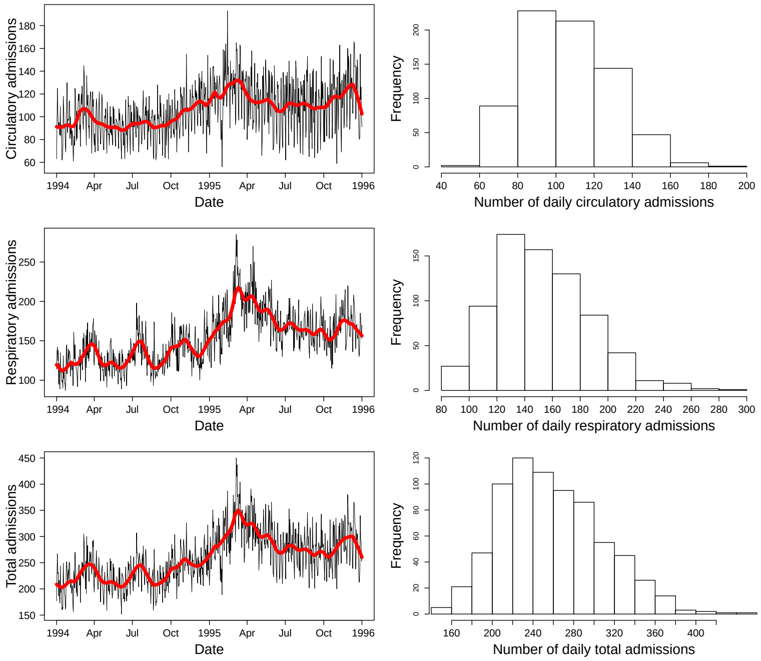


Figure 20: Daily hospital admissions in HongKong for 1994-1995.
Figures on the left plot the recorded observations against the trend (red curve). Figures on the right are histograms of daily hospital admissions.

In general, more people in Hong Kong suffered from respiratory diseases than circulatory problems in the two-year period, with mean daily hospital admissions of 106.2 and 153.2, respectively. And more hospital admissions were witnessed in 1995 than 1994 for both health problems. The risk of having circulatory diseases tended to be higher in winter and lower in summer times, showing a seasonal trend. Changing pattens in daily respiratory admissions were more complex. In 1994, the daily respiratory admission fluctuated and stayed in a relatively low level. It increased rapidly from the autumn of 1994 to the following spring. Before dropping and fluctuating to the level of approximately 150 patients per day, daily respiratory hospital admissions peaked in March 1995 at more than 250 patients daily.

10.1.2 Ambient air pollution

Sulphur dioxide (SO_2) - In an urbanised city, like Hong Kong, SO_2 is mainly contributed by usage of fossil fuels containing sulphur. More specifically, vehicles and industrial emissions in urban areas are major sources of SO_2 emission. Exposure to high levels of SO_2 risks damaging the functioning of respiratory system. For patients with respiratory and cardiac problems, high concentrations of SO_2 may aggravate their symptoms. Due to government interventions in controlling air pollution, the SO_2 concentrations have been maintaining at very low levels in Hong Kong. However, prolonged exposure at lower levels of SO_2 may also increase the risk of developing chronic respiratory disease.

Nitrogen dioxide (NO_2) - NO_2, which has potential to cause respiratory diseases is formed by oxidation of nitric oxide. Hong Kong has fairly high levels of NO_2 emission from power plants and diesel vehicles. Due to the proximity to the people, diesel vehicles are considered a more important risk factor to the functioning of respiratory system. The risk is higher in roadside areas and under calm wind conditions.

Respirable suspended particulates (RSP) - RSP (or PM_{10}) are particulate matters with aerodynamic diameters of 10 micrometres or smaller. Combustion processes and industrial emissions are significant sources of RSP in Hong Kong. RSP can penetrate deep into the lungs and cause chronic and acute pulmonary diseases. Its adverse effects on human health can be more risky if high RSP levels and higher levels of other pollutants coexist (Environmental Protection Department, 1996).

Ozone (O_3) - Ozone in this section means the ground-level ozone (or tropospheric ozone), rather than the Ozone layer in Earth's atmosphere. It is a sort of air pollution created near the Earth's surface by a series of complicated photochemical reactions under daylight UV rays. The combustion of fossil fuels produces major sources for the reactions. High levels of ozone can irritate eye, nose and throat and normal lung function. Prolonged exposure can aggravate respiratory infections.

	Min	25th Percentile	Medium	Mean	75th Percentile	Max
SO_2 ($\mu g/m^3$)	2.74	12.45	17.14	20.40	25.10	99.00
NO_2 ($\mu g/m^3$)	16.41	39.97	51.40	53.67	66.44	122.40
PM_{10} ($\mu g/m^3$)	14.77	30.73	45.25	50.58	66.28	159.70
O_3 ($\mu g/m^3$)	0.00	11.85	24.25	29.46	44.22	129.90

Table 30: Summary statistics of daily pollutant levels.

Thanks to the governmental lead control efforts, the SO_2 concentration levels in Hong Kong were maintained at very low levels (Environmental Protection Department, 1996,1997). Mean daily SO_2 concentration was only 20.4 g/m^3, which was lower than major western cities. However, the mean daily concentrations of NO_2 and PM_{10} were quite high, at 51.4 g/m^3 and

45.25 g/m^3, respectively. Ozone (O_3) concentrations were comparable to western cities (mean daily concentration at 29.46 g/m^3).

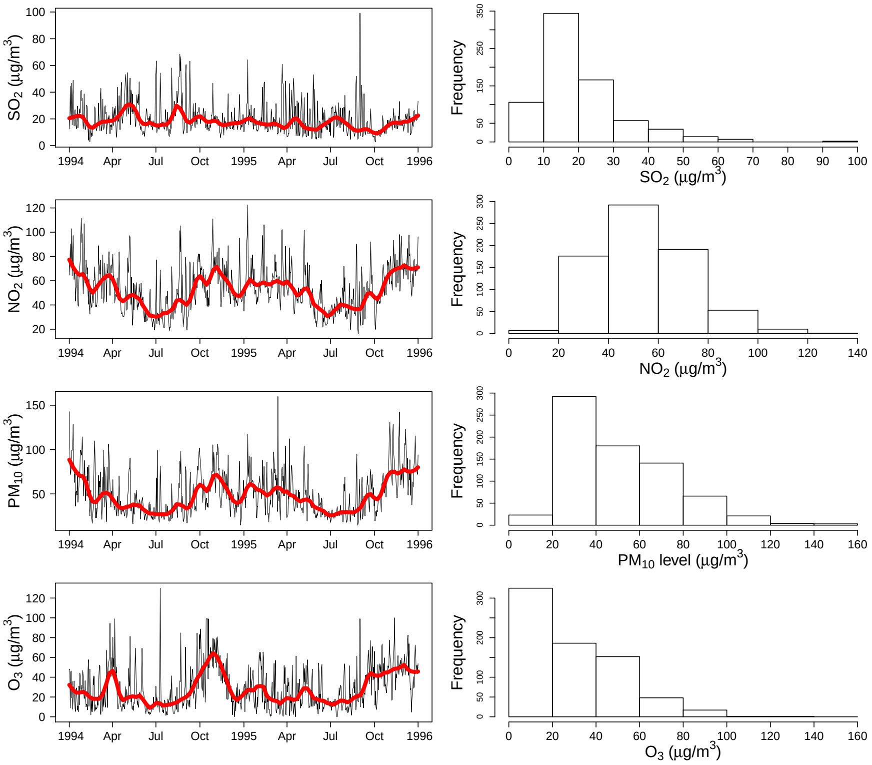


Figure 21: Daily air pollution levels in Hong Kong for 1994-1995.
Figures on the left plots the recorded daily average concentration of air pollutants against the trend (red curve). Figures on the right are histograms of daily average air pollution levels.

According to reports of Environmental Protection Department of Hong Kong (1997, 1998), air pollutant concentrations are generally lower in summer times due to the washout effect of rainfall and dispersion function of wind. This conclusion may explain well the changing patterns of NO_2 and PM_{10} emissions in Hong Kong for 1994 and 1995. The air pollution levels for NO_2 and PM_{10} were substantially lower in summer times. This is because of the washout effects of rainfall and better dispersion of pollutant during the summer months. Higher concentrations of air pollutants were, in general, recorded in the winter times, when the pollutants were trapped in Hong Kong due to weather conditions.

The patterns for SO_2 and ozone varied slightly from NO_2 and PM_{10}. SO_2 concentration levels did not show significant variation in the two-year period. The Environmental Protection Department (1996,1997) of Hong Kong attributes this phenomenon to higher electricity demand under hot weather conditions. In summer times, although the washout and dispersion effects were more dramatic than in winter days, they were offset by increased electricity consumption. For ozone, higher average concentrations occurred in October and November. The production of ozone is significant under strong day light, slow wind speed and low humidity conditions. October and November in 1994 and 1995 had more suitable weather conditions for photochemical formation of ozone.

10.2 Impact of air pollutants on health - respiratory and circulatory problems combined

Air pollutants are important factors that cause respiratory and circulatory problems. In their attempt of exploring the association between air pollution and health, Cai, Fan and Li (2000) combined these two groups of health problems and set the response variable as total number of daily hospital admissions for respiratory and circulatory problems. They constructed a GVCM with *log* transformation link function as

$$log\,[m(\mathbf{X}_i, t_i)] = a_0(t_i) + a_1(t_i)X_{1,i} + a_2(t_i)X_{2,i} + a_3(t_i)X_{3,i}, \qquad (10.1)$$

where X_1, X_2 and X_3 represent daily average concentration levels of SO_2, NO_2 and PM_{10}, and $m(\mathbf{X}_i, t_i) = E(y_i|\mathbf{X}_i, t_i)$ is the mean regression function of y given covariates $\mathbf{X}$ at time t. In case of confusion, note here that $a_0(\cdot)$ is the intercept term.

Since air pollutants are causal factors of circulatory and respiratory problems, it is reasonable to assume that the link function is strictly increasing. This makes the application of maximum rank correlation estimation method possible. In this thesis, we construct the model with monotonic link transformation. Moreover, the impact of ground-level ozone concentrations (X_4) is also taken into the model. Suppose that the conditional mean regression function $m(Y_i|\mathbf{X}_i, t_i)$, through an unknown link transformation, is linear as

$$g\Big(m(\mathbf{X}_i, t_i)\Big) = \beta_1(t_i)X_{1,i} + \beta_2(t_i)X_{2,i} + \beta_3(t_i)X_{3,i} + \beta_4(t_i)X_{4,i}, \qquad (10.2)$$

where $\boldsymbol{\beta}(\cdot) = \{\beta_1(\cdot), \beta_2(\cdot), \beta_3(\cdot), \beta_4(\cdot)\}^T$ is the varying coefficient vector of interest, and $g(\cdot)$ is the unknown strictly increasing link function.

Denote the directions of the varying coefficients as $\boldsymbol{\beta}_0(\cdot)$, the norm of the varying coefficients as $N(\cdot) = ||\boldsymbol{\beta}(\cdot)||$ and the first order derivative of the norm as $\dot{N}(\cdot) = c(\cdot)$. In our initial proposal, to make the model 10.2 identifiable, we commonly set $N(t_1) = ||\boldsymbol{\beta}(t_1)|| = 1$. However, the estimation of the varying coefficients at the beginning of 1994 frequently faces technical difficulties. These difficulties and their solutions will be demonstrated and introduced later. Instead of setting $N(t_1) = ||\boldsymbol{\beta}(t_1)|| = 1$, $N(t_{366}) = ||\boldsymbol{\beta}(t_{366})|| = 1$ is used as an alternative.

The directions of the varying coefficients are estimated with two-step estimation method. We do not known in advance where the true $c(\cdot)$ lies. To let the grid regression method make sense, at any date t, we search the estimator of $c(t)$ in a relatively large interval $c(t) \in [-60, 60]$. In terms of the selection of tuning parameters, it is necessary to use relatively large δ, since smaller δ would not allow the Newton-Raphson maximization algorithm produce converged maximizers. The tuning parameters are chosen to be $\delta = 0.3$ and $\lambda = 0.1$. Bandwidths are provided by cross-validation criterion regarding the objective function

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{m}_{-i}(y_i|\mathbf{X}_i, t_i))}{\sqrt{\hat{m}_{-i}(y_i|\mathbf{X}_i, t_i)}} \right)^2, \tag{10.3}$$

which is the mean value of the sum of squares of the Pearson's residuals, and where $\hat{m}_{-i}(\mathbf{X}_i, t_i))$ is the estimated conditional mean regression function of y at $t = t_i$ with the ith observations being deleted from the data. h denotes the set of bandwidths ($h_{20},\ h_{2j}, j = 1, \cdots, 4,\ h_n$, and h_l) applicable in the estimation, where h_{20} is the first stage bandwidth for two-step estimation of $\boldsymbol{\beta}(\cdot)$, $h_{2j}, j = 1, \cdots, 4$, are second stage bandwidths for two-step estimation of $\boldsymbol{\beta}(\cdot)$, h_n is the bandwidth for estimation of $||\boldsymbol{\beta}(\cdot)||$ and h_l is the link function estimation bandwidth.

- **Technical difficulty**

On constructing the model, ideally the varying coefficients, their norms and the first order derivatives of their norms are continuous and smooth functions. Originated from the feature of MRCE method proposed, estimation of $N(\cdot) = ||\boldsymbol{\beta}(\cdot)||$ is in fact approximation of $c(\cdot) = \frac{\dot{N}(\cdot)}{N(\cdot)}$. However, the estimation of $c(\cdot)$ could be extremely problematic.

At any t, the estimator $\hat{c}(t)$ of $c(t)$ is searched in the interval $[-60, 60]$. However, un-negligible amount of estimates reach the upper or lower bound of the searching region. Plot **A** in Figure 22 shows what the problems are like. It is vital to have a reasonable solution to this problem.

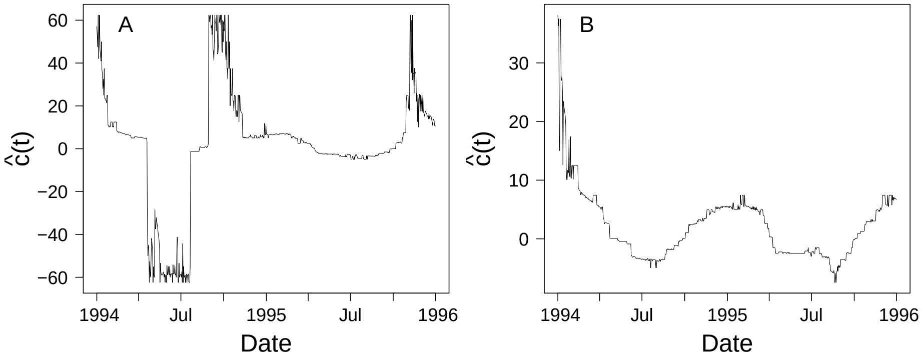


Figure 22: Technical difficulty and its solution.
(**A**) Problematic maximization of objective local rank correlation function. (**B**): Solution to problems in (**A**).

• Solution to technical issues

To conquer problematic maximization issue depicted in Plot **A** of Figure 22, the counting of local rank correlation scores shall be constrained. Suppose $\hat{\boldsymbol{\beta}}_0(\cdot)$ have being obtained. At any date t, denote $c(t) = \frac{\dot{N(t)}}{N(t)}$, $\mathbf{Z}_i = \mathbf{X}_i^T \hat{\boldsymbol{\beta}}_0(t_i)$ and $R = \{||\mathbf{Z}_i(1 + c(t_i - t))||, ||\mathbf{Z}_j(1 + c(t_j - t))||\}_{max}$, $i, j = 1 \cdots, n$. An objective function alternative to (3.11) is defined as

$$L(c) = \sum_{i \neq j} I(y_i > y_j) I \left(0 < \frac{\mathbf{Z}_i(1 + c(t_i - t)) - \mathbf{Z}_j(1 + c(t_j - t))}{R} \leq M \right),$$

where $M > 0$ is a constant. Maximization of this objective function at through t produces smooth maximizers of $c(\cdot)$. The selection of M should ensure that the objective rank correlation function is not overly constrained. It can be chosen by the data itself. However, the computation cost would dramatically be increased. For the data explored, $M = 10$ is used and practically useful.

$M = 10$ eliminates most of the wearied estimates of $c(\cdot)$. Unluckily, there are still a few poor estimates of $c(\cdot)$ remaining. As these poor estimates are minor, they are treated as outliers. When one obtains the estimates of $c(\cdot)$ at all datum points, say $\hat{c}(t_i), i = 1, \cdots, n$. Sort $\hat{c}(\cdot)$ in an increasing order and denote the new set as $c_o(t_i), i = 1, \cdots, n$. Estimates that are between 2.5% and 97.5% quantile of $c_o(\cdot)$ are retained, and the remaining 5% estimates are replaced by their local linear approximations. Plot **B** in Figure 22 shows that, with a proper constrain M, reasonable estimates of $c(\cdot)$ are obtained.

- **Results**

Minimization of cross validation defined in formula (10.3) gives bandwidths: $h_1 = 0.189$, $h_{21} = 0.126$, $h_{22} = 0.245$, $h_{23} = 0.358$, $h_{24} = 0.635$, $h_n = 0.189$ and $h_l = 0.177$. With the tricks proposed, the estimated functional coefficients are presented in Figure 23.

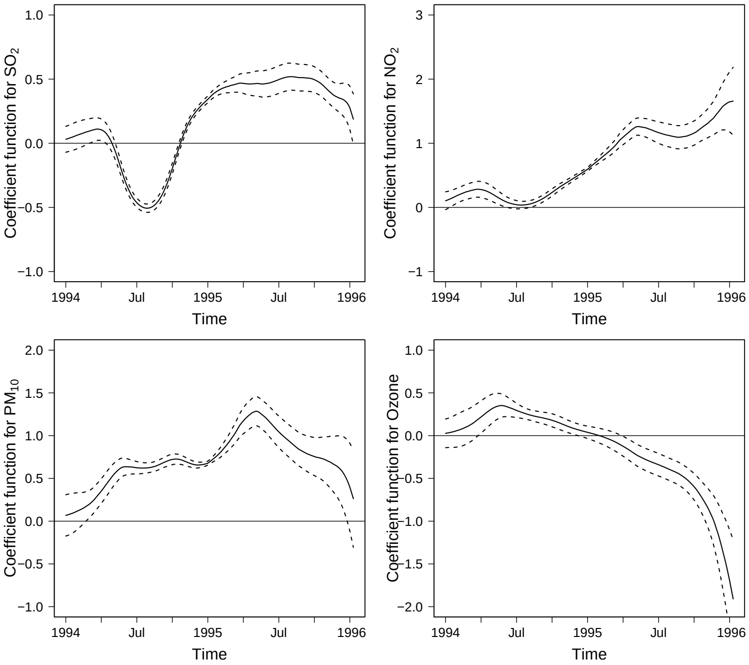


Figure 23: Estimated varying coefficient functions.
The solid lines are estimated coefficient functions, and the dashed lines are the estimated coefficient functions plus/minus twice estimated standard errors.

$\hat{\beta}_1(t)$ was predominantly negative before the winter of 1994, when its value became positive and climbed rapidly. The growth of $\hat{\beta}_1(t)$ slowed down in 1995, and $\hat{\beta}_1(t)$ began to drop from the level of 0.5 in September of 1995. $\hat{\beta}_2(t)$ and $\hat{\beta}_3(t)$ were positive in the two-year time. $\hat{\beta}_2(t)$ went up steadily between the summer of 1994 and the Spring of 1995 and stayed at a high level compared with other coefficients. $\hat{\beta}_3(t)$ increased rapidly in the Spring of 1994 and vibrated at the level just above 0.5. When it came to 1995, it experienced another sharp increase in the Spring and dropped gradually afterwards. The pattern of $\hat{\beta}_4(t)$ began with a gradual growth in the first four months of 1994. It gently decreased for more than one year's time and continued to decline sharply.

In general, major positive associations were found between daily hospital admissions for circulatory and respiratory diseases and daily average concentrations of two air pollutants - NO_2 and PM_{10}. Among the four air pollutants considered, NO_2(mean daily concentration at 53.67 g/m^3) and PM_{10}(mean daily concentration at 50.58 g/m^3) concentration levels were much higher in 1994 and 1995. Due to their dominance in ambient air pollution, for Hong Kong in 1994 and 1995, NO_2 and PM_{10} pollution were important explanatory factors to the emergence of circulatory and respiratory problems. Moreover, the changing patterns of $\hat{\beta}_2(t)$ and $\hat{\beta}_3(t)$ suggests that the risk from NO_2 tended to be more dominant than PM_{10}.

Harmful impact of SO_2 levels on health were only revealed in 1995, and the impact was smaller compared with NO_2 and PM_{10}. It is not surprising to see that the impact of SO_2 on the health problems in this study was not significant. Because, Hong Kong had controlled SO_2 emission efficiently and kept its concentrations to very low levels. What is surprising is that ground-level ozone tended to be negatively associated with daily hospital admissions for circulatory and respiratory problems. This might be explained by Pönkä and Virtanen (1996) who suggested that ozone is related to respiratory diseases but not to circulatory diseases.

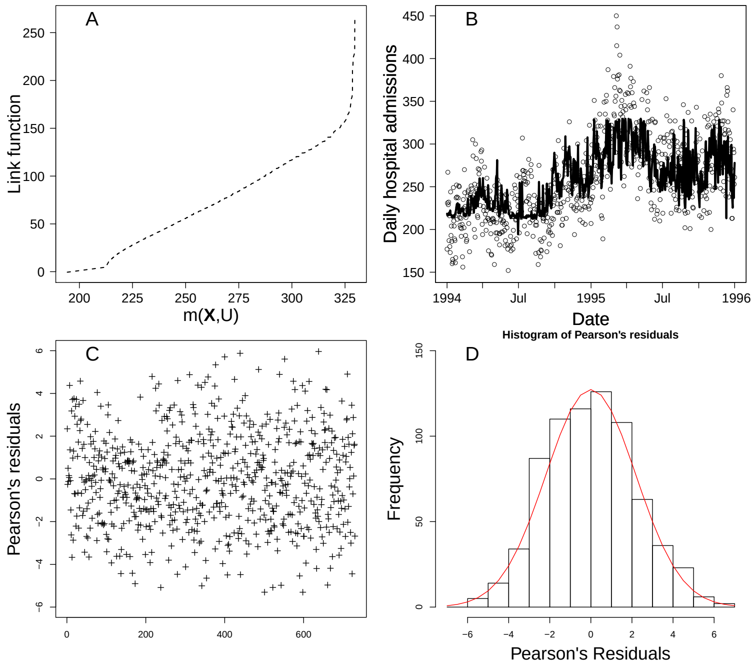


Figure 24: Link function and Pearson's residuals.
(**A**) Estimated curve of the unknown monotonic link function. (**B**) The solid line plots the estimated $\hat{m}(\mathbf{X}, t)$, where the dots are the observations of total daily hospital admissions for Circulatory and Respiratory problems. (**C**) Pearson's Residuals. (**D**) Histogram of Pearson's Residuals.

The proposed maximum rank correlation estimation method fits the data quite well. The Pearson's residuals are visually normal, indicating a proper approach to the mean regression function. The top left graph in Figure 23 shows the estimated link function. Despite the boundary where the estimates of the link function could be very bad, the estimated link function is visually linear. At least, it is neither the commonly used *log* function, nor any simple transformation of the *log* function.

This finding is interesting and attributes the proposed MRCE method more power in exploring real world problems. When it comes to count data, researchers commonly get an insight into the linear association between covariates and the response variable with some known link transformation. It is no doubt that the utilization of certain forms of link transformation is useful and meaningful. However, once the link function is very mistakenly defined, the model could be extremely wrong and providing misleading knowledge.

10.3 Short-term effects of air pollutants on respiratory and circulatory problems

The MRCE method has explored the association between daily hospital admissions for respiratory and circulatory problems and same-day ambient air pollution levels. What also interests us is the short-term effects of air pollution on health. Such short-term effects can be explored through the association between present hospital admissions and air conditions from a previous time to the present via

$$g\Big(m(\mathbf{X}_i, t_i, l)\Big) = \sum_{p=1}^{4} \beta_p(t_i) \Big(\sum_{j=0}^{l} X_{p,i-j} w_j \Big), \tag{10.4}$$

where l is the time lag, and w_j refers to the weight function of air pollutant level on lagged time t_{i-j}, $j = 0, \cdots, l$, with $\sum_{j=0}^{l} w_j = 1$. When $l = 0$, the model reduces to (10.2) in previous section.

For simplicity, when time lag is set to be l, identical weights, $w_j = \frac{1}{l+1}$, $j = 0, \cdots, l$, are used. Then the model consider short-term effects of air pollution on health is constructed as

$$g\Big(m(\mathbf{X}_i, t_i, l)\Big) = \sum_{p=1}^{4} \beta_p(t_i)\Big(\sum_{j=0}^{l} \frac{X_{p,i-j}}{l+1}\Big). \tag{10.5}$$

When time lag l is determined, of interest is to estimate the varying coefficients $\boldsymbol{\beta}(\cdot)$. The bandwidths are chosen by the CV defined as

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{m}_{-i}(\mathbf{X}_i, t_i, l))}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}}\right)^2. \tag{10.6}$$

The question is how to determine the time lag l. Due to that the interest is the short-term effects of air pollution on health, only time lag $l \leq 3$ are considered. For each value of l, standard sum of squares of Pearson's residuals (Criterion 1)

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{m}(\mathbf{X}_i, t_i, l)}{\sqrt{\hat{m}(\mathbf{X}_i, t_i, l)}}\right)^2, \tag{10.7}$$

and mean standard Pearson's residuals (Criterion 2)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{||y_i - \hat{m}_{-i}(\mathbf{X}_i, t_i, l))||}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}} \tag{10.8}$$

are calculated for time lag selection.

With $l = 2$, both criterion 1 (score = 4.68) and criterion 2 (score = 2.09) achieve the minimal values. Therefore, $l = 2$ is selected as the practical time lag. The CV criterion provides bandwidths: $h_1 = 0.228$, $h_{21} = 0.138$, $h_{22} = 0.222$, $h_{23} = 0.138$, $h_{24} = 0.326$, $h_n = 0.156$ and $h_l = 0.1$, which give the minimal cross validation score defined in formula (10.6). The estimated functional coefficients are presented in Figure 25.

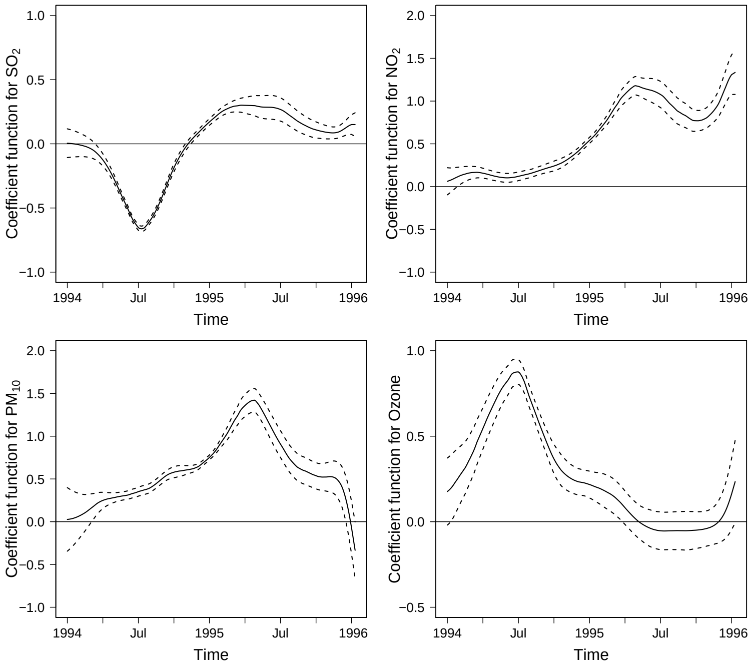


Figure 25: Estimated varying coefficient functions with time lag $l = 2$.
The solid lines are estimated coefficient functions, and the dashed lines are the estimated coefficient functions plus/minus twice estimated standard errors.

The patterns of $\hat{\beta}_1(\cdot)$, $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_3(\cdot)$ revealed by MRCE method with time lag $l = 2$ are similar to those given by $l = 0$ in previous section. Interestingly, the estimated curve for $\hat{\beta}_4(\cdot)$ is very different. $\hat{\beta}_4(\cdot)$ increased rapidly in the first half of 1994 and decreased fast in the following half-year time. In 1995, $\hat{\beta}_4(\cdot)$ was mainly positive. However, the values of $\hat{\beta}_4(\cdot)$ were close to 0.

With two days time lag, short-term effects of ground level ozone concentration levels were detected. Importantly, in 1994, ozone was a more significant air pollutant that other air pollutants that harmed human health. Because of low SO_2 levels in Hong Kong, SO_2 was still an insignificant factor that caused increasing hospital admissions for circulatory and respiratory diseases. Despite that the impact of PM$_{10}$ had decreased from April of 1995, positive associations between daily hospital admissions for circulatory and respiratory diseases and major air pollutants (NO_2 and PM$_{10}$) were dramatic in Hong Kong during the observatory period.

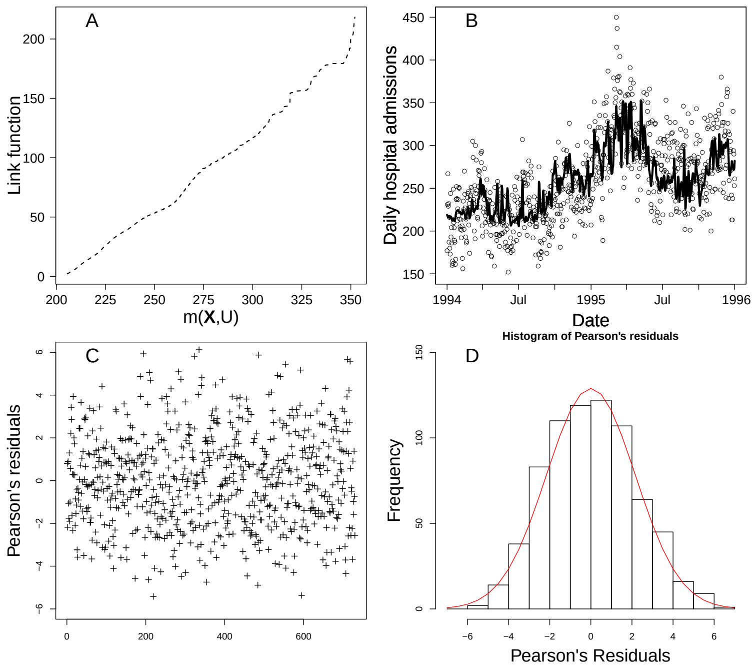


Figure 26: Link function and Pearson's residuals with time lag $l = 2$.
(**A**) Estimated curve of the unknown monotonic link function. (**B**) The solid line plots the estimated $\hat{m}(\mathbf{X}, t)$, where the dots are the observations of total daily hospital admissions for Circulatory and Respiratory problems. (**C**) Pearson's Residuals. (**D**) Histogram of Pearson's Residuals.

Normal Pearson's residuals indicates that the mean regression function is well estimated and the data set is reasonably fitted. Shown in plot **A** of 26, the estimated link function $g(\cdot)$ is increasing and close to linear.It is confident to conclude that the underline link function is not a *log* function. Therefore, the proposed MRCE method which assumes unknown monotonic link function is potential in explaining the data set more efficiently.

10.4 Two groups of diseases. Two Stories?

In previous analysis, the response variable is taken to be total number of daily hospital admissions of patients suffering from circulatory and respiratory problems. Under this setting, we actually assume identical link functions that associates the covariates with different groups of diseases. It is clear that air pollutants harm human bodies and cause circulatory and respiratory problems through diverse mechanisms. Therefore, in the model setting stage, it is more reasonable to take into consideration the potential that circulatory and respiratory diseases are related to air pollution in different ways. In this sense, link functions relate to these two health problems ought to be different. With different monotonic link functions, the two health problems are studied separately.

10.4.1 Short-term effects of air pollution on circulatory system

Suppose the conditional mean regression function $m(\mathbf{X}_i, t_i, l)$ is linear through an unknown strictly increasing link transformation $g_C(\cdot)$

$$g_C\Big(m(\mathbf{X}_i, t_i, l)\Big) = \sum_{p=1}^{4} \beta_p(t_i) \sum_{j=0}^{l} \Big(X_{p,i-j} w_j\Big), \tag{10.9}$$

where $\beta_p(\cdot)$, $p = 1, \cdots, 4$, are the varying coefficients, l is the time lag, and $w_j = \frac{1}{l+1}$, $j = 0, \cdots, l$, are the weights. To make the model identifiable, $N(t_{366}) = ||\boldsymbol{\beta}(t_{366})|| = 1$ is used.

The directions $\boldsymbol{\beta}_0(\cdot)$ of $\boldsymbol{\beta}(\cdot)$ are estimated with two-step estimation method. The estimator $\hat{c}(\cdot)$ of $c(\cdot) = \frac{\dot{N}(\cdot)}{N(\cdot)}$ at any t is searched in a relatively large interval [-60,60]. Acceptable bandwidths are provided by the CV criterion regarding the objective function

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{C,i} - \hat{m}_{-i}(\mathbf{X}_i, t_i, l)}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}}\right)^2. \tag{10.10}$$

Due to that the interest is the short-term effects of air pollution on health, only time lag $l \leq 3$ are considered. For each value of l, standard sum of squares of Pearson's residuals (Criterion 1)

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_{C,i} - \hat{m}(\mathbf{X}_i, t_i, l)}{\sqrt{\hat{m}(\mathbf{X}_i, t_i, l)}}\right)^2, \tag{10.11}$$

and mean standard Pearson's residuals (Criterion 2)

$$\frac{1}{n}\sum_{i=1}^{n}\frac{||y_{C,i} - \hat{m}_{-i}(\mathbf{X}_i, t_i, l))||}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}} \tag{10.12}$$

are calculated for time lag selection.

Evaluation of Criteria 1 and 2 suggests time lag $l = 0$ which achieved minimal value for Criterion 1 (scores at 3.54) and second smallest value for Criterion 2 (scores at 1.67). With time lag $l = 0$, the CV criterion suggests bandwidths: $h_1 = 0.251$, $h_{21} = 0.929$, $h_{22} = 0.222$, $h_{23} = 0.167$, $h_{24} = 0.698$, $h_n = 0.08$ and $h_l = 0.556$, which give the minimal cross validation score defined in formula (10.6).

As is depicted in Figure 27, $\hat{\beta}_1(\cdot)$, $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_3(\cdot)$ all went up gradually from January 1994 to September 1995. In the winter of 1995, $\hat{\beta}_1(\cdot)$ and $\hat{\beta}_2(\cdot)$ rocketed suddenly and, whereas $\hat{\beta}_3(\cdot)$ steeply dropped. The Pattern for $\hat{\beta}_4(\cdot)$ was very different form other functional coefficients. It was negative in the two-year period.

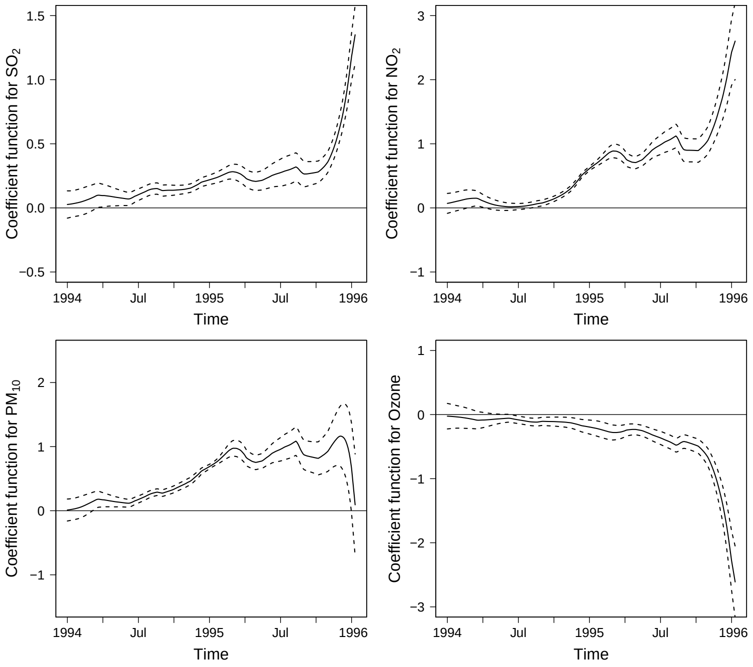


Figure 27: Estimated varying coefficient functions with time lag $l = 0$.
The solid lines are estimated coefficient functions, and the dashed lines are the estimated coefficient functions plus/minus twice estimated standard errors.

Major positive associations between circulatory diseases and air pollutants (NO_2 and PM_{10}) were detected in Hong Kong for 1994 and 1995. Although Hong Kong had low SO_2 levels, the MRCE method as well identified harmful effects of SO_2 on the circulatory system. In terms of impacts of ground level ozone on the circulatory system, our finding is in accordance with the claim of Pönkä and Virtanen (1996). Throughout 1994 and 1995, ground-level ozone is not directly related to circulatory diseases. The identified time lag $l = 0$ suggests that the impacts of air pollution on the circulatory system tended to be immediate and emergency for Hong Kong patients in 1994 and 1995. This finding advises the citizens to avoid too much roadside activities, because vehicles are major sources of air pollutants.

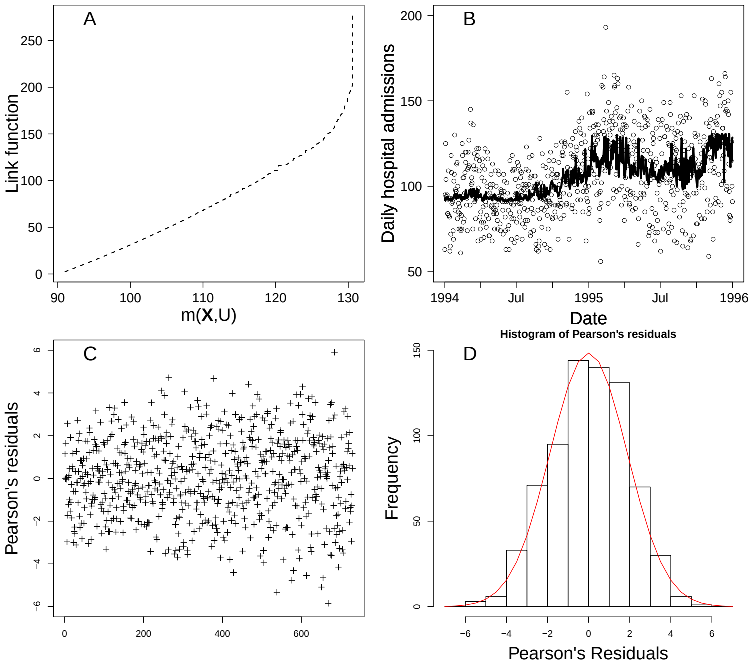


Figure 28: Link function and Pearson's residuals with time lag $l = 0$.
(**A**) Estimated curve of the unknown monotonic link function. (**B**) The solid line plots the estimated $\hat{m}(\mathbf{X}, t)$, where the dots are the observations of total daily hospital admissions for Circulatory and Respiratory problems. (**C**) Pearson's Residuals. (**D**) Histogram of Pearson's Residuals.

Normal Pearson's residuals indicates that the mean regression function is well estimated and the data set is reasonably fitted.Plot **A** of 26 depicts the estimated monotonic link function $g_C(\cdot)$. $g_C(\cdot)$ is not the commonly used *log* function. This confirms that approaching the data set with unknown link function is more practically realistic.

10.4.2 Short-term effects of air pollution on respiratory system

Suppose the conditional mean regression function $m(\mathbf{X}_i, t_i, l)$ is linear through an unknown strictly increasing link transformation $g_R(\cdot)$

$$g_R\Big(m(\mathbf{X}_i, t_i, l)\Big) = \sum_{p=1}^{4} \beta_p(t_i) \sum_{j=0}^{l} \Big(X_{p,i-j} w_j\Big), \tag{10.13}$$

where $\beta_p(\cdot)$, $p = 1, \cdots, 4$, are the varying coefficients, l is the time lag, and $w_j = \frac{1}{l+1}$, $j = 0, \cdots, l$, are the weights. To make the model identifiable, set $N(t_{366}) = ||\boldsymbol{\beta}(t_{366})|| = 1$.

The directions $\boldsymbol{\beta}_0(\cdot)$ of $\boldsymbol{\beta}(\cdot)$ are estimated with two-step estimation method. The estimator $\hat{c}(\cdot)$ of $c(\cdot) = \frac{\dot{N}(\cdot)}{N(\cdot)}$ at any t is searched in a relatively large interval [-60,60]. Reasonable bandwidths are provided by the CV criterion regarding the objective function

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_{R,i} - \hat{m}_{-i}(\mathbf{X}_i, t_i, l)}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}} \right)^2. \tag{10.14}$$

Due to that the interest is the short-term effects of air pollution on health, only time lag $l \leq 3$ are considered. For each value of l, standard sum of squares of Pearson's residuals (Criterion 1)

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_{R,i} - \hat{m}(\mathbf{X}_i, t_i, l)}{\sqrt{\hat{m}(\mathbf{X}_i, t_i, l)}} \right)^2, \tag{10.15}$$

and mean standard Pearson's residuals (Criterion 2)

$$\frac{1}{n} \sum_{i=1}^{n} \frac{||y_{R,i} - \hat{m}_{-i}(\mathbf{X}_i, t_i, l))||}{\sqrt{\hat{m}_{-i}(\mathbf{X}_i, t_i, l)}} \tag{10.16}$$

are calculated for time lag selection.

Both Criteria suggests time lag $l = 3$ which reached the minimal values for Criterion 1 (scores at 3.39) and Criterion 2 (scores at 1.69). With time lag $l = 3$, the CV criterion suggests bandwidths: $h_1 = 0.189$, $h_{21} = 0.222$, $h_{22} = 0.269$, $h_{23} = 0.222$, $h_{24} = 0.433$, $h_n = 0.156$ and $h_l = 0.214$, which give the minimal cross validation score defined in formula (10.6).

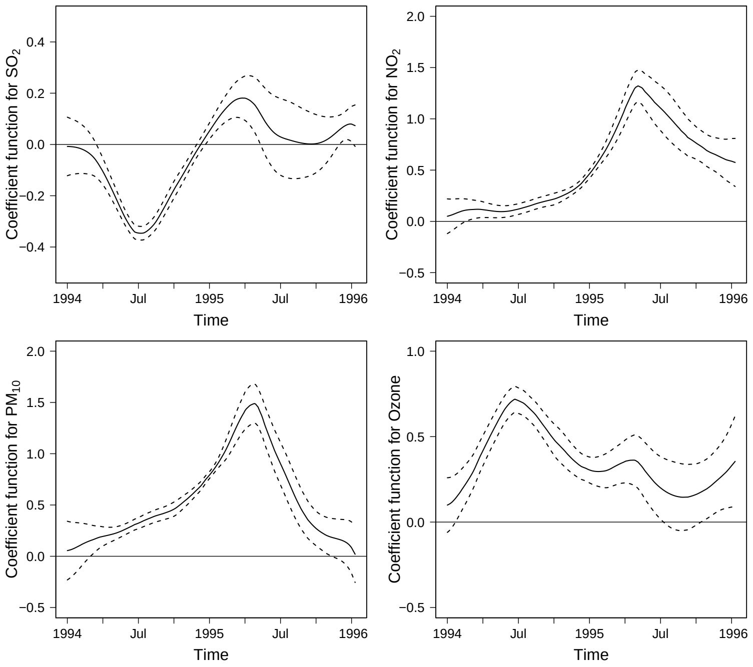


Figure 29: Estimated varying coefficient functions with time lag $l = 3$.
The solid lines are estimated coefficient functions, and the dashed lines are the estimated coefficient functions plus/minus twice estimated standard errors.

Depicted in Figure 29 are the estimated coefficient functions for ambient air pollutants. $\hat{\beta}_1(\cdot)$ was only positive in 1995, and its magnitude was much smaller than other coefficients. The changing patterns for positive functions $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_3(\cdot)$ were similar. Both coefficient functions had increased gradually in 1994 and then climbed quickly until May 1995. In the next eight months time, $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_3(\cdot)$ declined gradually, while $\hat{\beta}_3(\cdot)$ was dropping down faster than $\hat{\beta}_2(\cdot)$. $\hat{\beta}_4(\cdot)$ was also a positive function. It went up rapidly from January 1994 and peaked at the end of June that year. In the following one and a half year's time, $\hat{\beta}_4(\cdot)$ dropped gradually.

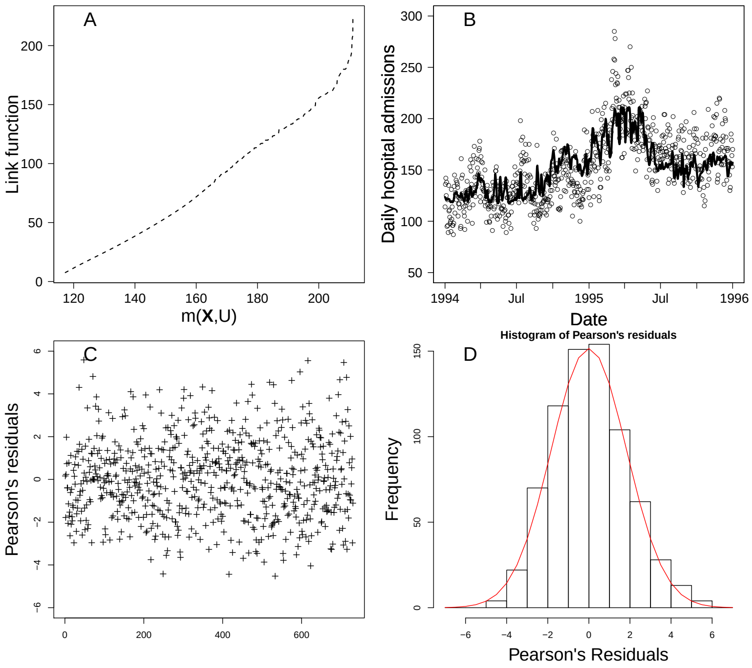


Figure 30: Link function and Pearson's residuals with time lag $l = 3$. (**A**) Estimated curve of the unknown monotonic link function. (**B**) The solid line plots the estimated $\hat{m}(\mathbf{X}, t)$, where the dots are the observations of total daily hospital admissions for Circulatory and Respiratory problems. (**C**) Pearson's Residuals. (**D**) Histogram of Pearson's Residuals.

It is found that respiratory diseases were positively associated with air pollutants (NO_2, PM$_{10}$ and ground-level ozone), while the impact of SO_2 on the respiratory system was not significant. Only very limited short-term effects of SO_2 were detected in 1995. High NO_2 and PM$_{10}$ concentration levels in Hong Kong were still important factors that risked the respiratory system, while positive association between ground level ozone and respiratory diseases was also revealed. In the study of circulatory diseases, impacts of air pollution on the circulatory system was severe ($l = 0$) in Hong Kong for 1994 and 1995. The situation was different for respiratory diseases in Hong Kong for the same time period. The impact of air pollution on the respiratory system required time to trigger paroxysm ($l = 3$). Normal Pearson's residuals confirms that the mean regression function is well estimated and the data set is reasonably fitted. The estimated monotonic link function $g_R(\cdot)$ depicted in Plot **A** of Figure 30 is close to linear rather than the commonly used *log* function.

Conclusion

Through the exploration of association between air pollution and health problems (respiratory and circulatory) in Hong Kong for 1994-1995, it is identified that harmful impact from SO_2 was limited, while NO_2 and PM_{10} were dramatic causal factors to both diseases. Ground-level was only related to respiratory diseases, whereas positive association with circulatory problems was not detected. The conducted analysis in this thesis is limited to situations in Hong Kong for 1994-1995 only. Conclusions can not be generalized Hong Kong at present time or other locations.

Additionally, the unknown strictly increasing link function was not commonly applied *log* function. This indicates that the proposed MRCE method for Generalized Varying Coefficient Models with unknown monotonic link function is more practically meaningful (when applicable) than methods that specify the link transformation.

Abbreviations

AIC - Akaike information criterion;

CV - Cross validation;

GLM - Generalized linear model;

GSVCM -Generalized semi-varying coefficient model;

GVCM - Generalized varying coefficient model;

i.i.d - Independent and identically distributed;

MSE - Mean squared error;

MISE - Mean integrated squared error;

MLE - Maximum likelihood estimation;

MRCE - Maximum rank correlation estimation;

VCM - Varying coefficient model;

WLS - Weighted least squares.

References

Barlow, R., Ireland, L., & Kass, L. (1982). Vision has a role in Limulus mating behaviour. *Nature*, **296(5852)**, 65-66.

Barlow, R., Powers, M., Howard, H., & Kass, L. (1987). Vision in Limulus mating and migration. In W. F. Herrnkind & A. B. Thistle, (Eds). *Signposts in the Sea.* Florida State University Press, Tallahassee, pp.69-84.

Borja-Aburto, V. H., Loomis, D. P., Bangdiwala, S. L., Shy, C. M., & Rascon-Pacheco, R. A. (1997), Ozone, suspended particulates, and daily mortality in Mexico City. *American journal of epidemiology*, **145(3)**, 258-268.

Brockmann, H. J. (1990). Mating behavior of horseshoe crabs, Limulus polyphemus. *Animal Behaviour*, **114(1)**, 206-220.

Brockmann, H. J., & Penn, D. (1992). Male mating tactics in the horseshoe crab, Limulus polyphemus. *Animal Behaviour*, **44(4)**, 653-665.

Brockmann, H. J. (1996). Satellite male groups in horseshoe crabs, Limulus polyphemus. *Ethology*, **102(1)**, 1-21.

Brunekreef, B., Dockery, D. W., & Krzyzanowski, M. (1995). Epidemiologic studies on short-term effects of low levels of major ambient air pollution components. *Environmental health perspectives*, **103(2)**, 3.

Cai, Z., Fan, J., and Li, R. (2000). Efficient Estimation and Inferences for Varying-Coefficient Models. *Journal of the American Statistical Association*, **95**, 888-902.

Cheng, M., Zhang, W., and Chen, L. (2009). Statistical Estimation in Generalized Multiparameter Likelihood Models. *Journal of the American Statistical Association*, **104**, 1179-1191

Cheng, S. (2002). Rank estimation of transformation models. *Econometrica*, **70**, 1683-1697.

Chiang, C. T., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, **96**, 605-619.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1991). Local regression models. In J. M. Chambers, and T. J. Hastie, (Eds). *Statistical Models in S*. Pacific Grove: Wadsworth & Brooks, pp.309-376.

Conte, S. D. and de Boor C. W. (1980). *Elementary Numerical Analysis: An Algorithmic Approach. 3rd edn.*. New York: McGraw-Hill.

Dab, W., Medina, S., Quenel, P., Le Moullec, Y., Le Tertre, A., Thelot, B., & Ferry, R. (1996). Short term respiratory health effects of ambient air pollution: results of the APHEA project in Paris. *Journal of epidemiology and community health*, **50(1)**, 42-46.

de Leon, A. P., Anderson, H. R., Bland, J. M., Strachan, D. P., & Bower, J. (1996). Effects of air pollution on daily hospital admissions for respiratory disease in London between 1987-88 and 1991-92. *Journal of Epidemiology and Community Health*, **50(1)**, 63-70.

Dobson, A.J. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

Dockery, D. W., Pope, C. A., Xu, X., Spengler, J. D., Ware, J. H., Fay, M. E., ... & Speizer, F. E. (1993). An association between air pollution and mortality in six US cities. *New England journal of medicine*, **329(24)**, 1753-1759.

Environmental Protection Department. (1997) Air quality in Hong Kong 1996. Hong Kong: Hong Kong Government, 1997:1-13.

Environmental Protection Department. (1998) Air quality in Hong Kong 1997. Hong Kong: Hong Kong Government, 1998:1-14.

Fan, J., and Huang, T. (2005), Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models, *Bernoulli*, **11**, 1031-1057.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, **27**, 1491-1518.

Fan, J., and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1(1)**, 179.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London

Faraway, J. J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *Journal of Statistical Computation and Simulation*, **37(1-2)**, 37-44.

Gilmour, P. S., Brown, D. M., Lindsay, T. G., Beswick, P. H., MacNee, W., & Donaldson, K. (1996). Adverse health effects of PM_{10} particles: involvement of iron in generation of hydroxyl radical. *Occupational and Environmental Medicine*, **53(12)**, 817-822.

Han, A. K. (1987). Non-parametric Analysis of a Generalized Regression Model. *Journal of Econometrics*, **35**, 303-316.

Hall, P. and Wilson, S. (1991). Two Guidelines for Bootstrap Hypothesis Testing. *Biometrics*, **47**, 757-762.

Hastie, T., and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*. **55**, 757-796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809-822.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variables. *Econometrica*, **64**, 103-137.

Huang, J. Z., Wu, C. O. and Zhou, L. (2002). Varying-coefficient mod- els and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111-128.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, **14**, 763-788.

Huang, J. Z. and Shen, H. (2004). Functional coefficient regression models for nonlinear time series: A polynomial spline approach. *Scandinavian Journal of Statistics*, **31**, 515-534.

Johnson, S. L., & Brockmann, H. J. (2012). Alternative reproductive tactics in female horseshoe crabs. *Behavioral Ecology*, **23(5)**, 999-1008.

Katsouyanni, K., Schwartz, J., Spix, C., Touloumi, G., Zmirou, D., Zanobetti, A., Wojtyniak, B., Vonk, J.M., Tobias, A., Pönkä, A., & Medina, S. (1996). Short term effects of air pollution on health: a European approach using epidemiologic time series data: the APHEA protocol. *Journal of Epidemiology and Community Health*, **50(1)**, 12-18.

Katsouyanni, K., Touloumi, G., Spix, C., Schwartz, J., Balducci, F., Medina, S., Rossi, G., Wojtyniak, B., Sunyer, J., Bacharova, L., & Schouten, J. P. (1997). Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project. *British Medical Journal*, **314(7095)**, 1658-1663.

Li, R., and Liang, H. (2008), Variable Selection in Semiparametric Regression Modeling. *Annals of Statistics*, **36**, 261-286

Lian, H. (2012). Variable selection for high-dimensional generalized varying-coefficient models. *Statistica Sinica*, **22**, 1563-1588

Lin, H and Peng, H. (2013). Smoothed rank correlation of the linear transformation regression model. *Computational Statistics and Data Analysis*, **57**, 615-630

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models, 2nd edn.* London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalised linear models. *Journal of the Royal Statistical Society: Series A*, **135**, 370 - 384.

Passaglia, C., Dodge, F. A., Herzog, E., Jackson, S. & Barlow, R. B. (1997). Deciphering a neural code for vision. *Proceedings of the National Academy of Sciences U.S.A.* **94**, 12649-12654.

Pönkä , A., and Virtanen, M. (1996). Low-level air pollution and hospital admissions for cardiac and cerebrovascular diseases in Helsinki. *American Journal of Public Health*, **86(9)**, 1273-1280.

Powers, M. K., Barlow, R. B., & Kass, L. (1991). Visual performance of horseshoe crabs day and night. *Visual Neuroscience*, **7**, 179-189.

Saldiva, P. H., Dockery, D. W., Pope, C. A., Schwartz, J., Dockery, D.W., Lichtenfels, A.J., Salge, J.M., Barone, I. and Bohm, G.M. (1995). Air pollution and mortality in elderly people: a time-series study in Sao Paulo, Brazil. *Archives of Environmental Health: An International Journal*, **50(2)**, 159-163.

Saunders, K. M., Brockmann, H. J., Watson, W. H., & Jury, S. H. (2010). Male horseshoe crabs *Limulus polyphemus* use multiple sensory cues to locate mates. *Current Zoology*, **56(5)**, 485-498.

Schouten, J. P., Vonk, J. M., de Graaf, A., (1996). Short term effects of air pollution on emergency hospital admissions for respiratory disease: results of the APHEA project in two major cities in The Netherlands, 1977-89. *Journal of epidemiology and community health*, **50(1)**,22-29.

Schwab, R. L. (2006). Mating group formation and female assessment by satellite male horseshoe crabs (Limulus polyphemus). *M.S. thesis, University of Florida, Gainesville, Florida.* Retrieved from http://etd.fcla.edu/UF/UFE0013901/schwab_r.pdf

Schwab, R. L., and Brockmann, H. J. (2007). The role of visual and chemical cues in the mating decisions of satellite male horseshoe crabs, Limulus polyphemus. *Animal Behaviour*, **74(4)**, 837-846.

Schwartz, J., Spix C., Touloumi G., Bachárová̌, L., Barumamdzadeh, T., le Tertre, A., Piekarksi, T., De Leon, A.P., Pönkä, A., Rossi, G. and Sáez, M. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *Journal of epidemiology and community health*, **50(1)**, 3-11.

Schwartz, J. (1996). Air pollution and hospital admissions for respiratory disease. *Epidemiology*, **7**, 20-28.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123-137.

Sunyer, J., Castellsagué, J., Sáez, M., Tobias, A., & Antó, J. M. (1996). Air pollution and mortality in Barcelona. *Journal of epidemiology and community health*, **50(1)**, 76-80.

Touloumi, G., Samoli, E., & Katsouyanni, K. (1996). Daily mortality and winter type air pollution in Athens, Greece: a time series analysis within the APHEA project. *Journal of epidemiology and community health*, **50(1)**, 47-51.

Vigotti, M. A., Rossi, G., Bisanti, L., Zanobetti, A., & Schwartz, J. (1996). Short term effects of urban air pollution on respiratory health in Milan, Italy, 1980-89. *Journal of epidemiology and community health*, **50(1)**, 71-75.

Wong, C. M., Ma, S., Hedley, A. J., & Lam, T. H. (1999) Does ozone have any effect on daily hospital admissions for circulatory diseases? *Journal of epidemiology and community health*, **53(9)**, 580-581.

Wong, T. W., Lau, T. S., Yu, T. S., Neller, A., Wong, S. L., Tam, W., & Pang, S. W. (1999). Air pollution and hospital admissions for respiratory and cardiovascular diseases in Hong Kong. *Occupational and environmental medicine*, **56(10)**, 679-683.

Wordley, J., Walters, S., & Ayres, J. G. (1997). Short term variations in hospital admissions and mortality and particulate air pollution. *Occupational and environmental medicine*, **54(2)**, 108-116.

Xu, X., Gao, J., Gao, J., & Chen, Y. (1994). Air pollution and daily mortality in residential areas of Beijing, China. *Archives of Environmental Health: An International Journalh*, **49(4)**, 216-222.

Ye, J.M., Duan, N.H. (1997). Nonparametric $n^{-1/2}$ consistent estimation for the general transformation models. *The Annals of Statistics*, **25**, 2682-2717.

Zhang, W. (2011). Identification of the constant components in generalised semivarying coefficient models by cross-validation. *Statistica Sinica*, **21**, 1913-1929.

Zhang, W., Lee, S. Y., & Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, **82(1)**, 166-188.

Zhang, W. and Peng, H. (2010). Simultaneous Confidence Band and Hypothesis Test in Generalised Varying-Coefficient Models. *Journal of Multivariate Analysis*, **101**, 1656-1680

Zhou, X. H., Lin, H., Johnson, E., (2008). Nonparametric heteroscedastic transformation regression models for skewed data with an application to health care costs. *Journal of the Royal Statistical Society: Series B*, **70(5)**, 1029-1047.

Zmirou, D., Barumandzadeh, T., Balducci, F., Ritter, P., Laham, G., & Ghilardi, J. P. (1996). Short term effects of air pollution on mortality in the city of Lyon, France, 1985-90. *Journal of epidemiology and community health*, **50(1)**, 30-35.