Regularized Statistical Methods for Data of Grouped or Dynamic Nature

by

Yun Li

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Naisyin Wang, Co-Chair
Associate Professor Ji Zhu, Co-Chair
Professor Kerby A. Shedden
Professor Peter Xuekun Song

THE UNIVERSITY OF MICHIGAN

ARTES · SCIENTIA · VERITAS

1817

To my mom, dad and lovely wife.

ACKNOWLEDGEMENTS

First of all, I would like to gratefully and sincerely thank my advisors, Professors Naiysin Wang and Ji Zhu, for their guidance throughout my Ph.D study. Without their tremendous help, this dissertation would not be finished smoothly. I also thank my other doctoral committee members, Professors Kerby Shedden and Peter Song, for their helpful comments and suggestions to this dissertation. Finally and most importantly, I would like to thank my mother, father and wife for their support, encouragement, quiet patience and unwavering love.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Regularized Statistical Methods for Data of Grouped or Dynamic Nature

by

Yun Li

Co-Chairs: Naisyin Wang and Ji Zhu

This dissertation consists of two parts. In the first part, one new convex regularized variable selection method is proposed for high-dimensional grouped data. Existing group variable selection methods via convex penalties, such as Yuan and Lin (2006) and Zhao et al. (2009), have the limitation of selecting variables in an "all-in-all-out" fashion and lack of selection flexibility within a group. In Chapter II, we propose a new group variable selection method via convex penalties that not only removes unimportant groups effectively, but also keeps the flexibility of selecting variables within an important group. Both the efficient numerical algorithm and high-dimensional theoretical estimation bounds are provided. Simulation results indicate that the proposed method works well in terms of both variable selection and prediction accuracy.

In the second part of the dissertation, we develop the parameter estimation methods for the dynamic ordinary differential equations (ODEs). Ramsay et al. (2007) proposed a popular parameter cascading method that tries to strike a balance between the data and the ODE structure via a "loss + penalty" framework. In Chapter III, we investigate this method in detail and take an alternative view through variance evaluation on it. We found, through both theoretical evaluation and numerical experiments, that the penalty term in Ramsay et al. (2007) could unnecessarily increase estimation variation. Consequently, we propose a simpler alternative structure for parameter cascading that achieves the minimum variation. We also provide theoretical explanations behind the observed phenomenon and report numerical findings on both simulations and one real dynamic data set.

In Chapter IV, we consider the estimation problem with time-varying ODE parameters. This is often necessary when there are unknown sources of disturbances that lead to deviations from the standard constant-parameter ODE system. To keep the structure of the parameters simple, we propose a novel regularization method for estimating time-varying ODE parameters. Our numerical studies suggest that the proposed approach works better than competing methods. We also provide finite-sample estimation error bounds under certain regularity conditions. The real-data applications of the proposed method lead to satisfactory and meaningful results.

CHAPTER I

Introduction and Literature Review

1.1 High-Dimensional Regularization Method

In many important statistical applications, the number of variables or parameters p is often much larger than the number of observations n. For example, the image data in radiology and biomedical science have only far fewer measurements of interest comparing with the unknown number of pixels in the images. Also, high-dimensional data often arise in genomics. In gene expression studies, due to the high cost of expression measurement, the number of observations is relatively low, typically, about tens to hundreds, while the total number of human gene assayed is commonly in thousands to ten thousands. This is the so-called large-p-small-n problem.

In statistical regression problems, it is of interest to recover the important predictors or the true signals relative to the response from the predictors with high-dimensional structure. In 1999, the Lasso method (Tibishirani, 1999) with the L_1 regularization penalty on the the loss function was proposed. Fan and Li (2001) provided the theoretical oracle properties for the Lasso method in the finite sample size (n is fixed) setting. The oracle properties when n is diverging was discussed in Fan and Peng (2004), but the p/n is still $o(1)$, and it is not for high-dimensional setting. Candes and Tao (2005) suggested the Danzig selector method and provided the high-dimensional properties for the corresponding method. They claimed that if the defined S-restricted isometry constant δ_S and S, S'-restricted orthogonality constants $\theta_{S,S'}$ of the data design matrix X satisfy certain inequalities, the estimation error bound between the estimated regression coefficient $\hat{\beta}$ and the true coefficient β is of the $\log p/n$ rate. That means that the Danzig selector works well when p grows almost at the exponential rate of n. To answer the question whether the Lasso method works well under the high-dimensional setting, Bickel et al. (2009) found that the Lasso method can provide the similar estimation error bound with the same rate of $\log p/n$ under certain eigenvalue restricted assumptions related to the design matrix X. They also pointed out the Danzig selector also works well for high-dimensional data under similar eigenvalue restricted assumptions of X and these assumptions are more easier to check comparing the restricted assumptions in Candes and Tao (2005). Simultaneously, Zhang and Huang (2008) obtained the same high-dimensional properties with the so-called sparse Riesz condition (SRC) of X. In addition to the estimation bounds, the high-dimensional sparse selection consistency for the the Lasso method was discussed in Meinshausen and Buhlmann (2006) and Meinshausen and Yu (2009) under some assumptions, such as, the irrepresentable condition.

In many situations, the conditional expectation of the response given the predictors may not be exactly linear. The Lasso and Danzig selector method for the linear regression needs to be extended to non-parametric regression. The non-parametric COSSO model was discussed in Lin and Zhang (2006). Meier et al. (2009) suggested one additive regression model and discussed the high-dimensional properties for it. The similar model, which is called SpAM, is also discussed in Ravikumar et al. (2008). In addition to statistical regression problems, the L_1 regularization method also applied to the graphical model (Yuan and Lin, 2007), and the high-dimensional properties of the precision matrix estimation were discussed in Lam and Fan (2009).

In many multiple regression problems predictors can be naturally grouped. For example, in the genomic studies, genes can be grouped with some certain pathway information. The group structures sometimes can help to find the important variables related to the response. The traditional Lasso method does not involve the group structure information. Instead of L_1 regularization penalty, Yuan and Lin (2006) proposed the group Lasso with the $\sqrt{L_2}$ penalty to penalize the loss function. The group Lasso method considers the group structure information in the penalty term, and performs better than the Lasso method for the grouped data. In Nardi and Rinaldo (2008), the high-dimensional properties for the group Lasso were discussed. In order to consider both group and within-group structure, Huang et al. (2009) and Zhou and Zhu (2010) proposed the $\sqrt{L_1}$ regularization method, which has superior performance in group and individual variable selection comparing the group Lasso and the Lasso for particular grouped data, which has both group and within-group structure. The high-dimensional theory has not yet been discussed for this regularization method. In Chapter 2, we design one new regularization method for the grouped data and discuss the high-dimensional properties for the corresponding method. The new method also can overcome some numerical issue comparing with the $\sqrt{L_1}$ regularization method.

1.2 Statistical Estimation Methods for Dynamic Systems

Most of the dynamic processes in engineering, biology and many other areas can often be modeled through ordinary differential equations (ODEs). For example, biologists use the FitzHugh-Nagumo ODE model (FitzHugh, 1961 and Nagumo et. al.,1962) to describe the behavior of spike potential in the giant axon of squid neurons. The Lotka-Volterra model (Lotka 1910 and Volterra 1926) is often used to model the dynamics of ecological systems with predator-prey interactions, competition, disease, and mutualism. Also, the Lotka-Volterra equations have a long history of use in economic theory. There exist parameters in the ODE models. Given the observation data contaminated with noises, how to estimate the parameters are of interest to both mathematicians and statisticians. Mathematicians developed the classical discretization schemes such as Euler or Runge-Kutta schemes in estimation algorithms based on the non-linear least squares method, for example, Hemker (1972) and Bard (1974). Statistician used the non-parametric estimation methods, for example, spline estimation methods (Varah, 1982 and Liang and Wu, 2008) and kernel estimation methods (Brunel, 2008), to estimate the ODE dynamic trajectories, and then the parameters are estimated followed by the non-linear least squares method. Recently, one promising statistical estimation method is proposed by Ramsay et al. (2008). In the so-called parameter cascading method, the function relationships between the ODE state components and system parameters are first constructed based on the basis function expansion or collocation methods. Then through applying the non-linear least squares method, one can precisely obtain the parameters of ODE dynamic systems.

In most of research works, the parameters are only considered as constants and do not change over time. But the dynamic systems are often delicate and sensible to the outer perturbations. For example, for the Lotka-Volterra model in the ecological studies, earthquakes or some unusual nature phenomena may break the balances and lead the parameters to change over time. Statistical estimation of time-varying parameters is very attracting. Chen and Wu (2008) applied the local polynomial estimation methods for the ODE trajectories and then estimated the time-varying parameters. Cao et al. (2011) applied the parameter cascading method with smoothing penalty to estimate the time-varying parameters. Actually the time-varying parameters sometimes have interpretable structures. For example, during some time the ODE system is balanced and the parameters keep constants over these time, and after some outer perturbation interferes, the balance is broken and then the parameters start to change over time. After a period of time, the ODE system may draw back to balance and then the parameters go back to constants again. Smoothing penalty may not work well to detect the structures of the time-varying parameters. Statistical estimation has not been discussed for the case when parameters vary over time with certain structures. We will discuss one regularization estimation methods in Chapter IV for this situation. Since the basis expansion method is applied in our estimation method, the high-dimensional regularization theory shows that the number of basis can be much larger than the number of observations in the obtained estimation error bounds under certain assumptions.

1.3 Organization of the Chapters

The dissertation is organized as follows. Chapter II studies the estimation problems for high-dimensional grouped data. One new regularization penalty is proposed and the complete algorithm for numerically solving the optimization problem is provided. We also demonstrate the high-dimensional properties for the corresponding method, and analyze one brain cancer real data using the new method. Chapter III and IV develop new statistical estimation methods for parameters in the dynamic systems controlled by ODEs. In Chapter III, we only consider the situation that the parameters of ODEs keep constants over time, and design one improved parameter cascading method to estimate the parameters and also the initial values of the ODE system. The estimation standard errors are reduced comparing with the existing method. One regularization method to estimate the time-varying parameter curves is discussed on Chapter IV. The estimation error bounds are obtained for both the parameter curves and the initial values. The asymptotic converge rates are also derived under certain conditions.

CHAPTER II

Convex Regularization Method for High-Dimensional Grouped Variable Selection

2.1 Introduction

Variable selection through optimizing a penalized log-likelihood function has been an active research area in the past decade. Most of the literature focuses on cases that the prediction variables do not form any group structure in a regression model. But in practice, the predictors are often naturally grouped. For example, in genomic studies, genes can be grouped in biological pathways that are related to phenotypes. Therefore it is of interest to study gene selection problems by taking into account group structures.

To address the variable selection problem when there are natural groups, Yuan and Lin (2006) proposed the group lasso method that penalizes the L_2-norm of the coefficients within each group, and Zhao et al. (2009) proposed to use the L_∞-norm penalty. Both methods are known to be able to remove unimportant groups effectively. One limitation of these two methods is that they select variables in an "all-in-all-out" fashion, i.e., when one variable in a group is selected, all other variables in the same group are also selected. This, however, may not be the case in practice, i.e., when a group of variables as a whole is important, the effects of some variables within this group may not be important. For example, in genomic studies, when a biological pathway is related to a certain phenotype, it does not necessarily mean that all genes in the pathway are related to the phenotype. Thus it is desirable to effectively remove unimportant groups while at the same time, be able to identify important variables within important groups.

To achieve this goal, Huang et al. (2009) proposed a group bridge approach, and Wang et al. (2009) and Zhou and Zhu (2010) independently investigated a hierarchical lasso approach that reduces to a special case of Huang et al. (2009). These methods can achieve the oracle property in the sense of Fan and Li (2001) and Fan and Peng (2004). Comparing with the methods by Yuan and Lin (2006) and Zhao et al. (2009), one drawback is that the objective functions are non-convex, which can cause numerical issues in practical computation; on a related matter, furthermore, the numerically obtained local optimum may not enjoy the theoretical optimal properties. To overcome this drawback, we propose a new group variable selection method in this paper. The new method has a convex objective function and can perform variable selection at both the group and within-group levels. We thoroughly investigate the new method both numerically and theoretically, and apply it to a glioblastoma microarray gene expression study conducted by Horvath et al. (2006).

This chapter is organized as follows. In Section 2.2, we propose the new regularization method and a corresponding algorithm. In Section 2.3, we develop non-asymptotic oracle inequalities in the high-dimensional setting where the number of prediction variables is allowed to be larger than the sample size. Simulation results are presented in Section 2.4. In Section 2.5, we apply the proposed method to a glioblastoma gene microarray study. Finally we conclude this chapter in Section 2.6.

2.2 Method

Consider a linear regression model with n observations:

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \ i = 1, \cdots, n, \tag{2.1}$$

where $X_{i1}, \cdots, X_{ip}$ are predictors, Y_i is the response variable, $\beta_1, \cdots, \beta_p$ are corresponding regression coefficients, and ε_i is the error term. We assume that all prediction variables and the response variable are centered, so we do not need to consider the intercept term in (2.1).

In this paper, we assume the prediction variables are naturally grouped. Specifically, we assume that the prediction variables can be divided into K groups, and the kth group contains p_k variables. Thus the linear model (2.1) can be rewritten as

$$Y_i = \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj} + \varepsilon_i, \ i = 1, \cdots, n, \tag{2.2}$$

where $X_{i,kj}$ denotes jth variable in the kth group of the ith observation.

To select variables at the group level, Yuan and Lin (2006) proposed the group lasso method that penalizes the L_2-norm of the coefficients within each group:

$$\min_{\beta_{kj}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj})^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k(\beta_{k1}^2 + \cdots + \beta_{kp_k}^2)}, \tag{2.3}$$

while Zhao et al. (2009) proposed to use the L_∞-norm penalty and minimize

$$\frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj})^2 + \lambda \sum_{k=1}^{K} \max\{|\beta_{k1}|, \cdots, |\beta_{kp_k}|\}. \tag{2.4}$$

Both methods are known to be able to remove unimportant groups effectively, but neither removes unimportant variables within an important group.

In order to also remove unimportant variables within important groups, Wang et al. (2009) and Zhou and Zhu (2010) considered a hierarchically penalized approach that reparametrizes $\beta_{kj} = \xi_k \theta_{kj}$ and minimizes

$$\min_{\xi_k, \theta_{kj}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \xi_k \theta_{kj} X_{i,kj})^2 + \lambda_\xi \sum_{k=1}^{K} \xi_k + \lambda_\theta \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\theta_{kj}|, \qquad (2.5)$$

where $\xi_k \geq 0$. It can be shown that the minimization problem in (2.5) is equivalent to the more general group bridge approach of Huang et al. (2009) with $\gamma = 0.5$:

$$\min_{\beta_{kj}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj})^2 + \lambda \sum_{k=1}^{K} (|\beta_{k1}| + \cdots + |\beta_{kp_k}|)^\gamma. \qquad (2.6)$$

In general, it is required $0 < \gamma < 1$.

Due to the singularity of the group bridge penalty, the above method is able to effectively remove unimportant groups. Furthermore, due to the singularity of the absolute value, the above method is also able to remove unimportant variables within identified important groups. However, one drawback of (2.6), compared with (2.3) and (2.4), is the non-convexity of the objective function in (2.6). This can cause a numerical issue in the sense that convergence to the global minimum is not guaranteed. Thus, theoretical optimal properties can not be guaranteed either for the numerically obtained local minimizer.

2.2.1 A convex penalty for grouped variable selection

To achieve the goal of both group and within-group variable selection and also to overcome the non-convex drawback, we propose a mixture of the weighted L_2-norm and L_1-norm penalties:

$$\min_{\beta_{kj}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj})^2 + \lambda_2 \sum_{k=1}^{K} \sqrt{\omega_k (\beta_{k1}^2 + \cdots + \beta_{kp_k}^2)} + \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}|, \qquad (2.7)$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters and ω_k is a weight coefficient. The L_2-norm and L_1-norm play different roles here: the L_2-norm penalty effectively removes unimportant groups while L_1-norm can remove unimportant variables within identified important groups. Furthermore, it can be easily verified that the objective function (2.7) is strictly convex.

To gain some insight with the method, we first show how it works under orthonormal design, i.e., $X^T X = I_p$. It is straightforward to obtain the following result, thus details are omitted.

Proposition II.1. *Assume that $X^T X = I_p$. Let X_{kj} denote the vector of length-n for the jth variable in the kth group. Let $\hat{\beta}_{kj}^{ols} = X_{kj}^T Y$ and $\hat{\beta}_{kj}^{lasso} = (\hat{\beta}_{kj}^{ols} - \lambda_1)_+ sgn(\hat{\beta}_{kj}^{ols})$ be the ordinary least square solution and lasso solution, respectively. Denote $\hat{\beta}_k^{ols} = (\hat{\beta}_{k1}^{ols}, \cdots, \hat{\beta}_{kp_k}^{ols})^T$ and $\hat{\beta}_k^{lasso} = (\hat{\beta}_{k1}^{lasso}, \cdots, \hat{\beta}_{kp_k}^{lasso})^T$. Let $\hat{\beta}_{kj}$ denote the solution of (2.7) and $\hat{\beta}_k = (\hat{\beta}_{k1}, \cdots, \hat{\beta}_{kp_k})^T$, then we have*

$$\hat{\beta}_k = \begin{cases} \left(1 - \frac{\lambda_2 \sqrt{\omega_k}}{\|\hat{\beta}_k^{lasso}\|_2}\right)_+ \hat{\beta}_k^{lasso}, & \textit{if } \hat{\beta}_k^{lasso} \neq 0, \\ 0, & \textit{if } \hat{\beta}_k^{lasso} = 0. \end{cases} \tag{2.8}$$

From the above proposition, we can see that the solution to (2.7) in the orthonormal setting can be viewed as being obtained by a two stage method. It first shrinks $\hat{\beta}_{kj}^{ols}$ by a soft-thresholding at λ_1. Then it further shrinks the weighted L_2-norm of the coefficient vector of each group by a soft-thresholding at λ_2. It turns out that unimportant groups are effectively removed in the second step, and some variables within identified important groups may have already been removed in the first step. From solution (2.8) we can also see that if $\omega_k = 1$ for all k, then since groups with small numbers of coefficients tend to have small $\|\hat{\beta}_k^{lasso}\|_2$, their coefficients tend to be shrunk to zero more easily relative to larger groups even if they are important groups. This is not a desirable feature. Therefore, ω_k can be chosen to compensate for that. In the rest of the paper, we choose $\omega_k = p_k$, the length of each group.

2.2.2 Algorithm

When the design matrix is not orthonormal, analytical solution of (2.7) in general does not exist. In this subsection, we develop a shooting algorithm (Fu 1998, Friedman et al. 2007) for solving (2.7). The shooting algorithm is essentially a "coordinate descent" algorithm. That is, in each iteration we fix all but one coefficient, say β_{kj}, at their current values, then optimize (2.7) to solve for β_{kj}. Since this optimization only involves one parameter, it is often easy to achieve a solution. The algorithm can be formulated as follows:

1. (Standardization) Standardize each variable such that

$$\sum_{i=1}^{n} Y_i = 0, \ \sum_{i=1}^{n} X_{i,kj} = 0, \ \text{ and } \ \sum_{i=1}^{n} X_{i,kj}^2 = 1. \tag{2.9}$$

2. (Initialization) Initialize $\beta_{kj}^{(0)}$ with some plausible values, and set $m = 1$.

3. (Update β_{kj}) Fix $\beta_{k'j'}$ at $\beta_{k'j'}^{(m-1)}$, $k' \neq k$ or $j' \neq j$, and solve for β_{kj}. With a little algebra, we obtain the following. If $\beta_{kj'}^{(m-1)} = 0$ for $j' \neq j$, then

$$\beta_{kj}^{(m)} = \frac{\left(|S_{kj}^{(m-1)}| - (\lambda_1 + \lambda_2\sqrt{p_k})\right)_+}{X_{kj}^T X_{kj}} \operatorname{sgn}(S_{kj}^{(m-1)}), \tag{2.10}$$

where $S_{kj}^{(m-1)} = X_{kj}^T(Y - X\beta_{-kj}^{(m-1)})$ and $\beta_{-kj}^{(m-1)}$ is the same as the coefficient vector $\beta^{(m-1)}$ except that the kjth element is equal to 0; else if $\beta_{kj'}^{(m-1)} \neq 0$ for some $j' \neq j$, then

$$\beta_{kj}^{(m)} = \frac{\left(|S_{kj}^{(m-1)}| - \lambda_1\right)_+}{X_{kj}^T X_{kj} + \lambda_2\sqrt{p_k}\left(\beta_{kj}^{(m)2} + \sum_{j' \neq j} \beta_{kj'}^{(m-1)2}\right)^{-1/2}} \operatorname{sgn}(S_{kj}^{(m-1)}). \tag{2.11}$$

Note that both sides of (2.11) involve $\beta_{kj}^{(m)}$, thus the solution $\beta_{kj}^{(m)}$ can be achieved by iterating between the two sides of (2.11).

4. If $\|\beta^{(m)} - \beta^{(m-1)}\|_2$ is less than a pre-specified tolerance value, then stop the algorithm. Otherwise, let $m \leftarrow m+1$ and go back to Step 3.

In practice, one often needs to optimize (2.7) over a large number of (λ_1, λ_2) pairs. One strategy of increasing computing efficiency is to start with very large (λ_1, λ_2) and initialize $\beta_{kj}^{(0)} = 0$, and gradually reduce the values of λ_1 and λ_2. When (λ_1, λ_2) become smaller, one can initialize $\beta_{kj}^{(0)}$ using $\hat{\beta}_{kj}$ at previous values of (λ_1, λ_2). In our numerical experiments, we found this is very effective in reducing the computational cost.

2.2.3 Tuning parameters

There are several commonly used tuning parameter selection methods, such as cross-validation, generalized cross-validation (GCV), AIC and BIC, where

$$\begin{aligned} \text{GCV} &= \frac{\|Y - X\hat{\beta}\|_2^2}{n(1 - df/n)^2}, \\ \text{AIC} &= \log(\|Y - X\hat{\beta}\|_2^2/n) + 2df/n, \\ \text{BIC} &= \log(\|Y - X\hat{\beta}\|_2^2/n) + \log n \cdot df/n. \end{aligned}$$

Yang (2005) noted that AIC and GCV are more suitable if the selected model is used for prediction, while for the purpose of variable selection, BIC is more appropriate. Note that the above criteria all depend on *df*, the degree of freedom of the selected model. Here we develop two ways to approximate *df*. The first one is based on Karuch-Kuhn-Tucker conditions, and the second one is based on the Stein's identity.

For the first one, let $\hat{\beta}$ be the minimizer of (2.7). The Karuch-Kuhn-Tucker conditions imply

$$X_{kj}^T(Y - X\hat{\beta}) = \lambda_2 \frac{\sqrt{p_k}\hat{\beta}_{kj}}{\sqrt{\sum_{j=1}^{p_k}\hat{\beta}_{kj}^2}} + \lambda_1 \text{sgn}(\hat{\beta}_{kj}), \ \forall \hat{\beta}_{kj} \neq 0. \tag{2.12}$$

Denote $\mathcal{A} = \{kj : \hat{\beta}_{kj} \neq 0\}$. Let $\hat{\beta}_{\mathcal{A}}$ be a vector containing non-zero $\hat{\beta}$ and $X_{\mathcal{A}}$ be the corresponding design matrix. Then since $\text{sgn}(\hat{\beta}_{kj}) = \hat{\beta}_{kj}/|\hat{\beta}_{kj}|$, we have

$$\hat{Y} = X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}} = X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}} + W)^{-1} X_{\mathcal{A}}^T Y, \tag{2.13}$$

where W is a diagonal matrix with elements:

$$\frac{\lambda_2\sqrt{p_k}}{\sqrt{\sum_{j=1}^{p_k}\hat{\beta}_{kj}^2}} + \frac{\lambda_1}{|\hat{\beta}_{kj}|}, \ kj \in \mathcal{A}.$$

Mimicking the ridge regression, the number of effective parameters, df, can be approximated by

$$df(\lambda_1, \lambda_2) = \text{Tr}(X_{\mathcal{A}}(X_{\mathcal{A}}^T X_{\mathcal{A}} + W)^{-1} X_{\mathcal{A}}^T). \tag{2.14}$$

This approximation is similar to that in Tibshirani (1996) and Fu (1998) for lasso.

For the second method, we consider the orthonormal design setting, i.e., $X^T X = I_p$, and employ the Stein identity $df = \sum_{i=1}^n \text{cov}(\hat{Y}_i, Y_i)/\sigma^2 = E(\sum_{i=1}^n \partial\hat{Y}_i/\partial Y_i)$. With a little algebra, we can obtain

$$\begin{aligned} df &= \sum_{k=1}^K I(\|\hat{\beta}_k^{lasso}\|_2 > \lambda_2\sqrt{p_k}) \\ &\quad + \sum_{k=1}^K I(\|\hat{\beta}_k^{lasso}\|_2 \neq 0)\left(1 - \frac{\lambda_2\sqrt{p_k}}{\|\hat{\beta}_k^{lasso}\|_2}\right)_+ \left(\sum_{j=1}^{p_k} I(\hat{\beta}_{kj}^{lasso} \neq 0) - 1\right) \end{aligned} \tag{2.15}$$

If $\lambda_2 = 0$, (2.15) reduces to $|\mathcal{A}|$, which is what is proposed in Zou et al. (2007). If

$\lambda_1 = 0$, (2.15) becomes

$$df = \sum_{k=1}^{K} I(\|\hat{\beta}_k^{ols}\|_2 > \lambda_2 \sqrt{p_k}) + \sum_{k=1}^{K} \left(1 - \frac{\lambda_2 \sqrt{p_k}}{\|\hat{\beta}_k^{ols}\|_2}\right)_+ (p_k - 1), \qquad (2.16)$$

which is the same as what is given in Yuan and Lin (2006).

Note that (2.14) tends to be smaller than (2.15). For example, when $\lambda_2 = 0$, df in (2.14) is less than or equal to $|\mathcal{A}|$. This implies that (2.14) may underestimate the effective number of parameters. Indeed in our simulation studies, we found that (2.15) performs better than (2.14).

2.3 Non-asymptotic properties

In this section, we study non-asymptotic properties of (2.7). We are interested in estimating methods that work well even when the number of predictors is much larger than the number of observations, i.e., $p \gg n$, since this situation may arise in many practical applications. Meinshausen and Buhlmann (2006), Zhang and Huang (2008) and Meinshausen and Yu (2009) showed that the lasso method works well under high dimensional settings. Bickel et al. (2009) proved that both the lasso and the Danzig selector have similar oracle bounds in the case $p \gg n$. Naidi and Rinaldo (2008) used the similar idea of Bickel et al. (2009) to show that the bounds of group lasso has the same rate as lasso if the tuning parameters for the two methods have the same rate. We will extend the argument of Bickel et al. (2009) to show that similar bounds also hold for our proposed method.

Denote the true coefficient vector as β^*. Suppose we have K groups and each group has the same number of coefficients L for notational simplicity. Therefore ω_k can be omitted. For every $\beta \in \mathbb{R}^p$, $p = KL$, we denote $\beta^k = (\beta_{k1}, \cdots, \beta_{kL})^T$, the coefficient vector for the k-th group. Denote

$$\|\beta\|_2 = \sum_{k=1}^{K} \|\beta^k\|_2, \quad |\beta|_1 = \sum_{k=1}^{K} |\beta^k| = \sum_{k=1}^{K}\sum_{j=1}^{L} |\beta_{kj}|.$$

We rewrite our problem as follows:

$$\min_{\beta_{kj}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{k=1}^{K}\sum_{j=1}^{L}\beta_{kj}X_{i,kj})^2 + 2\lambda_2\|\beta\|_2 + 2\lambda_1|\beta|_1. \tag{2.17}$$

Let

$$\begin{aligned} M(\beta) &= \sum_{kj,1\le k\le K,1\le j\le L} I(\beta_{kj}\neq 0) = |J(\beta)| \quad \text{and} \\ M_G(\beta) &= \sum_{k,1\le k\le K} I(\|\beta^k\|\neq 0) = |J_G(\beta)|, \end{aligned}$$

respectively denote the number of non-zero coefficients and the number of groups with a non-zero coefficient norm, where $I(\cdot)$ is the indicator function, $J(\beta) = \{kj : \beta_{kj} \neq 0\}$ and $J_G(\beta) = \{k : \|\beta^k\| \neq 0\}$. Both $|J|$ and $|J_G|$ denote the cardinalities of the corresponding sets. The value of $M(\beta)$ measures the sparsity of the coefficient vector β, and so does $M_G(\beta)$ at the group level.

For a vector $\Delta \in \mathbb{R}^p$ and a subset $J \subset \{1, \cdots, p\}$, we denote by Δ_J the vector in $\mathbb{R}^p$ with the same coordinates as Δ on J and zero coordinates on the complement J^c of J. For $J_G \subset \{1, \cdots, K\}$, we also denote by Δ_{J_G} the vector with the same coordinates as Δ on J_G groups and zero coordinates on the complement J_G^c of J_G.

Let $M(\beta^*) = r$ and $M_G(\beta^*) = s$. We make the following assumption that is similar to Bickel et al. (2009).

Assumption II.2. *RE(r, s, ρ): Assume K groups with L variables in each group. For a vector $\Delta \in \mathbb{R}^p$ satisfying $|\Delta_{J^c}| + 2\rho\|\Delta_{J_G^c}\|_2 \le 3|\Delta_J| + 2\rho\|\Delta_{J_G}\|_2$ for $|J| \le r$ and*

$|J_G| \leq s$, where $\rho \geq 0$, we assume

$$\begin{aligned} \kappa(r,s,\rho) &\triangleq \min_{J,J_G,\Delta\neq 0} \frac{\|X\Delta\|}{\sqrt{n}\|\Delta_J\|} > 0, \\ \kappa_G(r,s,\rho) &\triangleq \min_{J,J_G,\Delta\neq 0} \frac{\|X\Delta\|}{\sqrt{n}\|\Delta_{J_G}\|} > 0. \end{aligned} \tag{2.18}$$

Theorem II.3. *Consider model (2.2). Assume that the random variables $\varepsilon_1, \cdots, \varepsilon_n$ are independent and follow a normal distribution with mean zero and variance σ^2, and all diagonal elements of the matrix X^TX/n are equal to 1. Suppose $M_G(\beta^*) = s$ and $M(\beta^*) = r$. Furthermore, suppose Assumption RE(r,s,ρ) holds with $\kappa_G = \kappa_G(r,s,\rho)$ and $\kappa = \kappa(r,s,\rho)$, and let $\phi_{\max}$ be the largest eigenvalue of the matrix X^TX/n. Let*

$$\lambda_2 = \rho\lambda_1 = 2\rho A\sigma\sqrt{\frac{\log p}{n}},$$

and $A > \sqrt{2}$. Then with probability at least $1 - p^{1-A^2/2}$, for $\hat{\beta}$ that minimizes (2.17), we have

$$\begin{aligned} \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2 &\leq 64A^2\sigma^2 \cdot \frac{\log p}{n}\left(\frac{\rho\sqrt{s}}{\kappa_G} + \frac{\sqrt{r}}{\kappa}\right)^2, \qquad (2.19) \\ \|\hat{\beta}-\beta^*\|_2 &\leq \frac{32A\sigma}{2\rho+1} \cdot \sqrt{\frac{\log p}{n}}\left(\frac{\rho\sqrt{s}}{\kappa_G} + \frac{\sqrt{r}}{\kappa}\right)^2, \qquad (2.20) \\ M(\hat{\beta}) &\leq 64\phi_{\max}\left(\frac{\rho\sqrt{s}}{\kappa_G} + \frac{\sqrt{r}}{\kappa}\right)^2. \qquad (2.21) \end{aligned}$$

If Assumption RE$(2r, 2s, \rho)$ also holds, then with the same probability we have

$$\|\hat{\beta}-\beta^*\|_2 \leq \frac{32\sqrt{2}A\sigma}{2\rho\sqrt{s}+\sqrt{r}} \cdot \sqrt{\frac{\log p}{n}}\left(\frac{\rho\sqrt{s}}{\kappa_G(2r,2s,\rho)} + \frac{\sqrt{r}}{\kappa(2r,2s,\rho)}\right)^2. \tag{2.22}$$

The proof is given in Appendix A. The theorem tells us that the method (2.7) is rate-consistent in model selection. The rate is the same as the lasso rate (Bickel et al., 2009) and the group lasso rate (Naidi and Rinaldo, 2008). Therefore, even if the exact sparsity pattern may not be fully recovered, the estimator can still be a good approximation to the truth. This theorem also suggests that it might be easier to achieve the estimation consistency than the variable selection consistency.

Since our new method includes the lasso penalty term, the conditions are relaxed comparing with those in Naidi and Rinaldo (2008). In their proof for the group lasso method, a tail probability bound for the chi-square distribution with the degree of freedom equal to the length of group L is applied. This bound will diverge as the length of a group diverges. Thus in order to achieve convergence in probability, they needed the condition that L cannot diverge faster than $\log p/n$. This condition is relaxed in our theorem due to the lasso penalty in (2.7).

2.4 Simulation studies

In this section we compare our method (2.7) with lasso (Tibshirani, 1996) and group lasso (Yuan and Lin, 2006) using several simulation studies. To compensate for the possible over-shrinkage caused by the double-penalty in (2.7), following Zou and Hastie (2005), we propose to adjust $\hat{\beta}_{kj}$ after optimizing (2.7). Specifically, based on Proposition 2.1, we propose to adjust $\hat{\beta}_{kj}$ as follows:

$$\hat{\beta}_k^{adj.} = \begin{cases} \frac{\|\hat{\beta}_{kj}^{lasso}\|_2}{\|\hat{\beta}_k^{lasso}\|_2 - \lambda_2\sqrt{p_k}}\hat{\beta}_{kj}, & \text{if } \|\hat{\beta}_k^{lasso}\|_2 > \lambda_2\sqrt{p_k}, \\ \hat{\beta}_{kj}, & \text{otherwise.} \end{cases} \tag{2.23}$$

We consider two simulation set-ups. Each set-up includes two different cases: one is “all-in-all-out”, i.e., if a group is important all variables in the group are important, and the other is “not-all-in-all-out”. In Simulation I, all groups are of the same length, while in Simulation II, different groups may have different numbers of covariates.

Simulation I

There are 8 groups and each group consists of 6 covariates. We first generate independent random variables $R_{1,1}, \cdots, R_{1,6}, R_{2,1}, \cdots, R_{8,6}$ from $N(0,1)$ and correlated random variables $Z_1, \cdots, Z_8$ from $N(0,1)$ with an AR(1) covariance structure, i.e., $\text{Cov}(Z_k, Z_{k'}) = \rho^{|k-k'|}$ for $1 \leq k, k' \leq 8$. Then we construct the covariate vector $(X_{1,1}, \cdots, X_{8,6})$ as follows:

$$X_{kj} = (Z_k + R_{kj})/\sqrt{2},\ 1 \leq k \leq 8,\ 1 \leq j \leq 6.$$

Thus variables within the same group are correlated, and variables belonging to different groups are also correlated when $\rho \neq 0$. In the "all-in-all-out" case, we set β^* as

$$\beta^* = (\underbrace{1, 1.5, 2, 2.5, 3, 3.5}_{6}, \underbrace{2, 2, 2, 2, 2, 2}_{6}, \underbrace{0, \cdots, 0}_{6\times 6})^T,$$

and in the "not-all-in-all-out" case, we set β^* as

$$\beta^* = (\underbrace{1, 0, 2, 0, 3, 0}_{6}, \underbrace{2, 0, 2, 0, 2, 0}_{6}, \underbrace{0, \cdots, 0}_{6\times 6})^T.$$

Simulation II

There are also 8 groups. Each of the first four groups has 8 variables and each of the remaining four groups has 4 variables. Similar to Simulation I, we also generate independent random variables $R_{1,1}, \cdots, R_{8,4}$ from $N(0,1)$ and correlated random variables $Z_1, \cdots, Z_8$ from $N(0,1)$ with an $AR(1)$ structure, i.e., $\text{Cov}(Z_k, Z_{k'}) = \rho^{|k-k'|}$ for $1 \leq k, k' \leq 8$. Then we construct the covariates as follows:

$$X_{kj} = (Z_k + R_{kj})/\sqrt{2},\ 1 \leq k \leq 8,\ 1 \leq j \leq 8 \text{ for } k \leq 4,\ 1 \leq j \leq 4 \text{ for } k > 4.$$

In the "all-in-all-out" case, we set β^* as

$$\beta^* = (\underbrace{1, 1.5, 2, 2.5, 1, 1.5, 2, 2.5}_{8}, \underbrace{0, \cdots, 0}_{3\times 8}, \underbrace{1, 1, 2, 2}_{4}, \underbrace{0, \cdots, 0}_{3\times 4})^T,$$

and in the "not-all-in-all-out" case, we set β^* as

$$\beta^* = (\underbrace{1,0,2,0,1,0,2,0}_{8}, \underbrace{0,\cdots,0}_{3\times 8}, \underbrace{1,0,0,2}_{4}, \underbrace{0,\cdots,0}_{3\times 4})^T.$$

For each of the above settings, we consider two different values of ρ: $\rho = 0.5$ and 0.8, and we generate the response variable Y by

$$Y = \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta^*_{kj} X_{kj} + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. We set the value of σ such that the signal-to-noise ratio is equal to 1, which is quite low and mimics the real scenario.

We consider two training sample sizes: $n = 100$ and $n = 200$, and we apply three methods to select tuning parameters: GCV, AIC and BIC. For the final selected models, we compare the sensitivity and specificity that are defined as

$$\begin{aligned} \text{sensitivity} &= \frac{\#\{j : \hat{\beta}_j \neq 0 \text{ and } \beta^*_j \neq 0\}}{\#\{j : \beta^*_j \neq 0\}}, \\ \text{specificity} &= \frac{\#\{j : \hat{\beta}_j = 0 \text{ and } \beta^*_j = 0\}}{\#\{j : \beta^*_j = 0\}}. \end{aligned}$$

We also compute the prediction error (PE) on a separate test set with sample size 1000. We repeat 100 times for each setting, and report the average results in Figures 2.1-2.4.

We note that the GCV and AIC have similar performance, so we only report the results of GCV. We also notice that BIC tends to work better than GCV in terms of model selection, and GCV tends to work better than BIC in terms of prediction. Therefore, we only report the results of BIC for sensitivity and specificity, and the results of GCV in terms of prediction error.

Figures 2.1 and 2.2 show that the group lasso tends to achieve high specificity in the “all-in-all-out” case. This indicates that group lasso is effective in removing unimportant groups. Compared with group lasso, for which the setting of “all-in-all-out” is in favor of, our methods (both $L_1 + L_2$ and adjusted $L_1 + L_2$) sacrifice a little in terms of specificity, but gain a lot in sensitivity. The setting of “not-all-in-all-out”, however, is not in favor of group lasso, and our methods, especially adjusted $L_1 + L_2$, perform better than group lasso in terms of both sensitivity and specificity. The difference is more prominent in the setting where the sample size is relatively small ($n = 100$). Comparing the new methods with lasso, it seems that the new methods always dominate lasso in terms of both sensitivity and specificity in both settings.

Figures 2.3 and 2.4 show that our methods perform better than both lasso and group lasso in terms of the prediction error, where the adjusted $L_1 + L_2$ performs the best in most cases. The inferiority of group lasso is more prominent in Simulation II than in Simulation I.

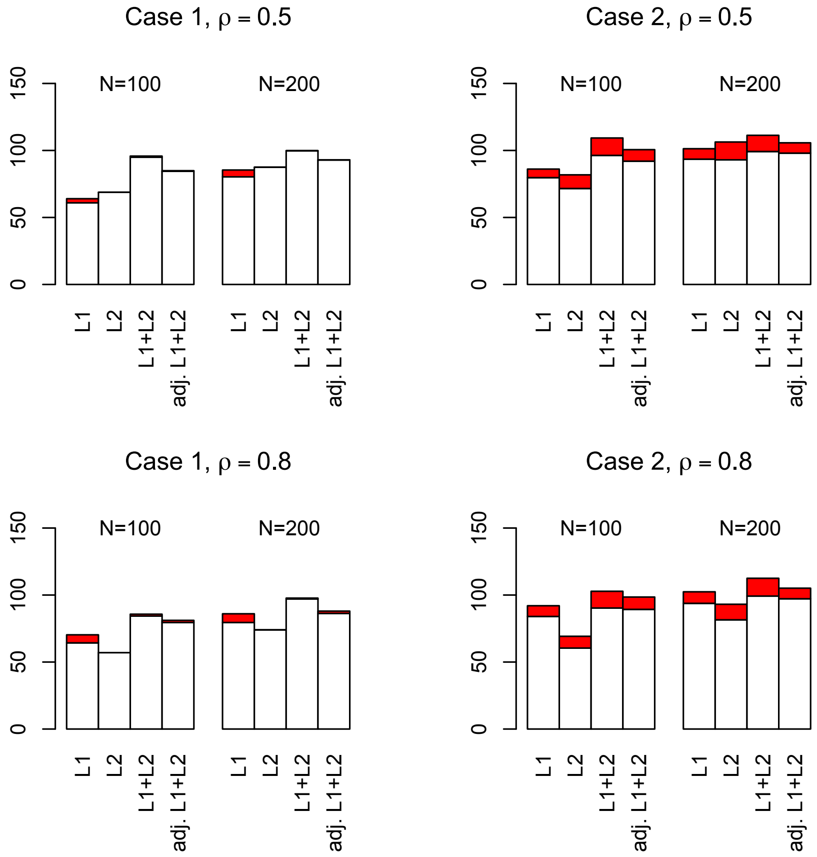

Figure 2.1: Simulation I. Comparison of variable selection. White box: sensitivity; red box: 1-specificity.

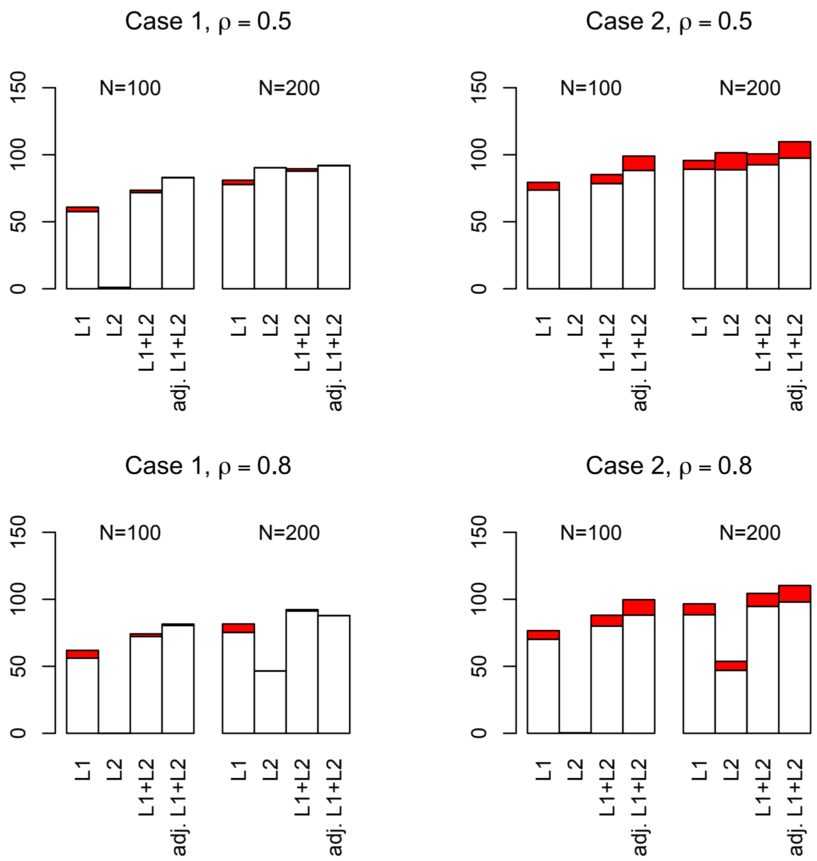

Figure 2.2: Simulation II. Comparison of variable selection. White box: sensitivity; red box: 1-specificity.

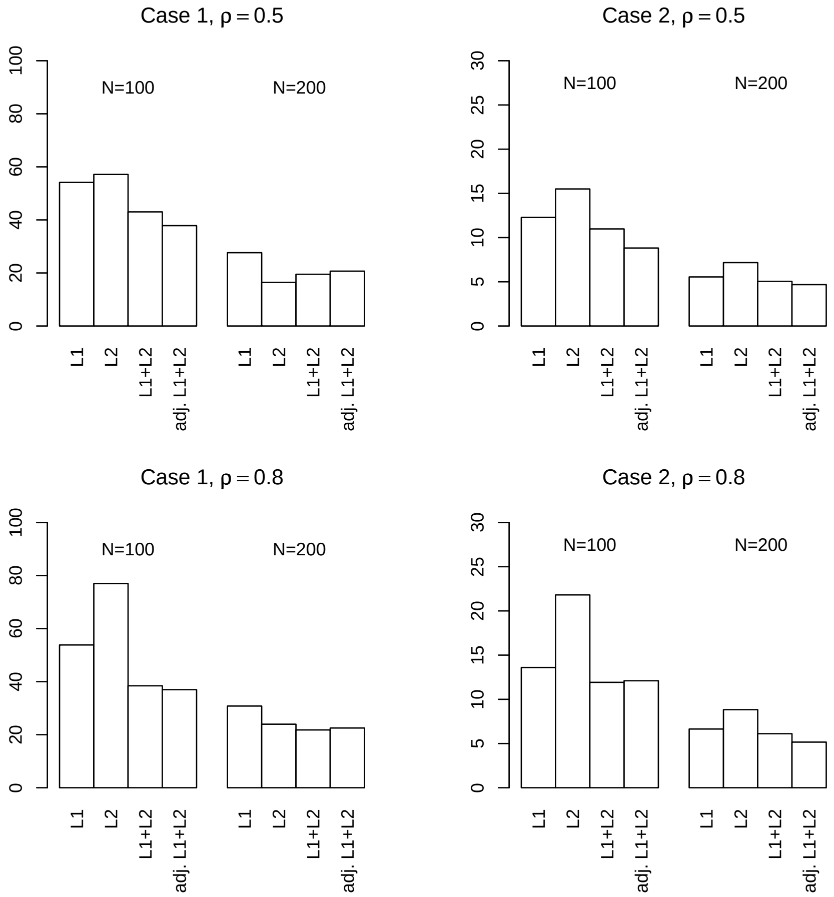

Figure 2.3: Simulation I. Comparison of prediction error.

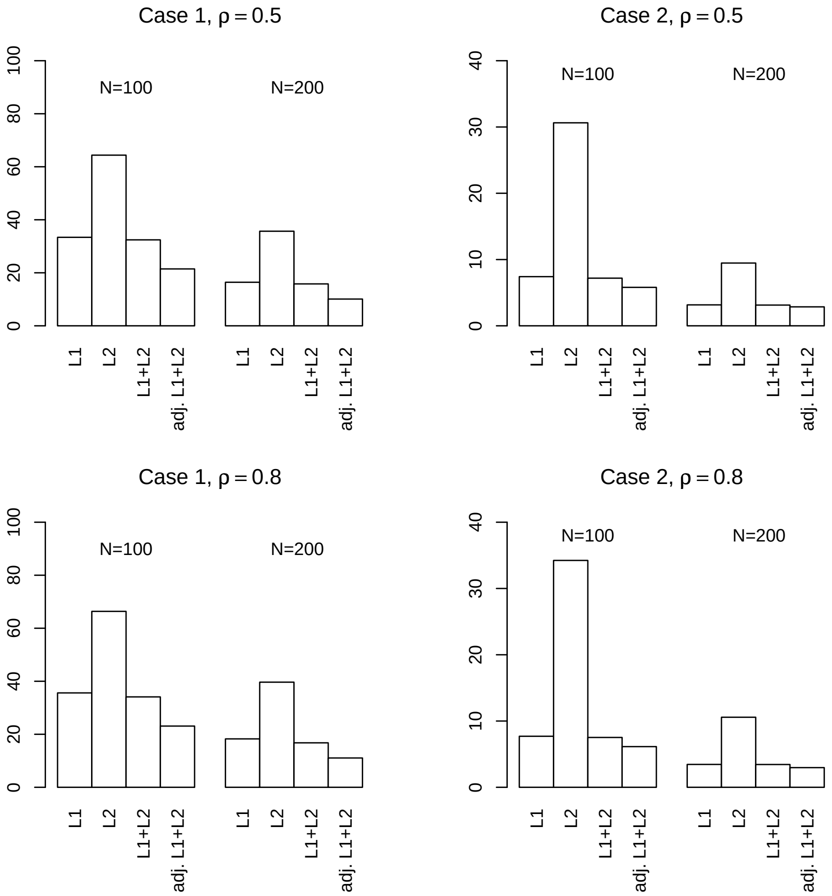

Figure 2.4: Simulation II. Comparison of prediction error.

2.5 Data example

Glioblastoma is the most common primary malignant brain tumor of adults and one of the most lethal of all cancers. Patients with this disease have a median survival of 15 months from the time of diagnosis despite surgery, radiation and chemotherapy. In this section, we apply our method to a glioblastoma microarray gene expression study conducted by Horvath et al. (2006). Global gene expression data from n=120 clinical tumor samples are obtained by high-density Affymetrix arrays, and each sample contains expression values of 8,000 genes. Among the 120 patients, 9 were alive at the last followup. In our analysis, these 9 censored subjects are excluded. We take the logarithm of survival time of 111 patients in days as the response variable.

Horvath et al. (2006) found 3,600 most connected genes from the original 8,000 genes and constructed a weighted brain cancer network. Starting with those 3,600 genes, we first assessed each of the 3,600 genes by running simple linear regression on the data set, and reduced the number of genes to 1,000 using the p-values. This type of univariate screening process was justified in theory by Fan and Lv (2008). We then used the gene group information on the GSEA (gene set enrichment analysis) website (http://www.broadinstitu- te.org/gsea/) to construct group structures among the genes. Two kinds of gene group sets were used in our analysis. One is the gene pathway (PW), and the other is the gene ontology (GO). The genes associated with the same PW or GO set are considered as one group. Among the 1,000 genes that passed the initial screening, we identified 799 genes in gene pathway groups and 508 genes in gene ontology groups.

One noticeable phenomenon is that some genes belong to two or more groups, i.e., different groups may have overlapping genes. We make a small modification to the penalty in (2.7) to accommodate for overlapping groups. Specifically, we consider the criterion

$$\min_{\beta_{kj}} \frac{1}{2} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} \beta_{kj} X_{i,kj})^2 + \lambda_2 \sum_{k=1}^{K} \sqrt{p_k \left(\frac{\beta_{k1}^2}{n_{k1}} + \cdots + \frac{\beta_{kp_k}^2}{n_{kp_k}} \right)} + \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\beta_{kj}| \tag{2.24}$$

where n_{kj} is the number of groups that contain the kj-th variable X_{kj}. The algorithm proposed in Section 2.2.2 can also be revised accordingly. The change is straightforward; we omit the details here.

We randomly split the samples into the training and test sets 100 times; for each split, two-thirds of the samples were used for training and the rest were for testing. For each split, we applied two methods, our proposed method and lasso. Tuning parameters were chosen using five-fold cross-validation.

We first compare the prediction accuracy of the two methods (Table 2.1). Over the 100 random splits, our proposed method has an average prediction error of 0.774 (when using PW groups) and 0.748 (when using GO groups), which is respectively smaller than 0.798 and 0.786 of lasso. This is consistent with what we have observed in simulation studies.

Next, we compare gene selection of the two methods. From Table 2.2, we can see that when using GO groups (the top half of Table 2.2), 12 genes were selected more than 80 times out of the 100 random splits by our method, while 10 genes were selected by lasso, and the 10 genes selected by lasso are a subset of the 12 genes selected by our method. The results are similar when we use other thresholds, such as 70, 60 and 50 for the number of times of a gene being selected, i.e., our method tends to select more genes than lasso and the genes selected by lasso tend to be a subset of the genes selected by our method. The pattern remains the same when the PW groups are used (the bottom half of Table 2.2).

We have also looked at the individual genes that were selected in Table 2.1. For example, both our method and lasso selected CSF1R with high frequency (more than

80%). It is known that CSF1R is expressed in glioblastoma cell-lines and in operative specimens of human glioblastoma (Alterman and Stanley, 1994). There are also genes that were selected by our method with high frequency but not by lasso, while in the literature believed to be associated with glioblastoma, e.g. GPNMB and GADD45B. Kuan et al. (2006) and Tse et al. (2006) pointed out that GPNMB is a cell-surface proteoglycan expressed by melanoma and glioblastoma cells. Le Mercier et al. (2008) found that decreasing Galectin-1 expression, which also impairs the expression level of GADD45B, in human orthotopic glioblastoma xenografts significantly increases the survival time of glioblastoma tumor-bearing mice.

These results indicate that our method is more powerful than lasso in identifying important genes, which is again consistent with what we have observed in simulation studies.

Group structure	Our method	Lasso
Gene pathway	0.774(0.267)	0.798(0.276)
Gene ontology	0.748(0.189)	0.786(0.202)

Table 2.1: Comparison of prediction error with two different group structures

Gene ontology			
Threshold	Our method	Lasso	Common genes
80%	12	10	10
70%	26	18	17
60%	53	25	25
50%	78	50	48
Gene pathway			
Threshold	Our method	Lasso	Common genes
80%	8	8	6
70%	24	18	16
60%	57	35	30
50%	126	63	62

Table 2.2: Comparison of the number of selected genes with different selection frequency threshold

2.6 Summary

In this chapter, we have proposed a new method with convex penalties for group variable selection. The new method keeps the advantage of group lasso in terms of effectively removing unimportant groups, while at the same time also enjoys the flexibility of removing unimportant variables within identified important groups. We have developed an efficient shooting algorithm for solving the corresponding optimization problem, and we have also established non-asymptotic error bounds for the new method, in which the number of prediction variables is allowed to be much larger than the sample size. Numerical results indicate that the proposed new method works well in terms of both prediction accuracy and variable selection, especially when variables in a group are associated with the response in a “not-all-in-all-out” fashion.

CHAPTER III

Parameter Estimation for Ordinary Differential Equations

3.1 Introduction

In physics, engineering, economics and biological sciences, dynamics systems are often described by ordinary differential equations (ODEs). We consider an ODE model,

$$\frac{dX(t)}{dt} = F\{X(t), \theta\},$$

where $X(t) = \{X_1(t), \cdots, X_m(t)\}^{\mathrm{T}}$ is the true unobserved state vector to describe the dynamic system, $\theta = (\theta_1, \cdots, \theta_d)^{\mathrm{T}}$ denotes the unknown parameters to be estimated, and $F(\cdot) = \{F_1(\cdot), \cdots, F_m(\cdot)\}^{\mathrm{T}}$ gives a known functional structure. In practice, not all $X_i(t)$ can be observed and the ones that can be observed may be observed with an additive error $\varepsilon_i(t)$. Consequently, we model the observed $Y(t)$ as

$$Y_i(t) = X_i(t) + \varepsilon_i(t), \ \ \text{for } i = 1, \ \cdots, \ m_o,$$

where $m_o \leq m$ and m_o indicates the number of observable components; $\varepsilon_i(t)$ is assumed to be independent identically distributed mean-zero error at time t. Note that we do not assume the errors are correlated here. One scenario that we will discuss in details is the model mis-specification. It would be difficult to distinguish between two possibilities: the model structures being mis-specified versus part of the errors are of time series nature.

It is important to precisely estimate parameters θ in a dynamics system. There is a long history in parameter estimation for ODE models. Hemker (1972) and Bard (1974) rely on the non-linear least squares (NLS) method and the standard numerical ODE solver, such as the Runge-Kutta algorithm, estimating θ^*. Varah (1982) further adopts the spline smoothing techique. Recently, Ramsay et al. (2007) proposed a promising strategy for ODE parameter estimation based on a "collocation" method and penalized NLS algorithm. Hereafter, we refer to this nominal work as RHCC. RHCC proposed a parameter-cascades and profiling procedure that estimates the ODE curves via basis function expansions and spline smoothing. Their method enables the use of analytical gradient and Hessian in the NLS procedure. The authors incorporated a penalized optimization procedure within their parameter-cascades estimation method. Even though the tuning parameter(s) selection could be computationally burdensome, the authors aim to use this step to strike a balance between the goodness of fit to the observed data and fidelity to the ODE structure.

Liang and Wu (2008) adopted a local polynomial smoothing approach to the estimation of ODE curves and their corresponding first derivatives and proposed an alternative. They plugged-in the estimates into the optimization procedure to obtain the estimates of ODE parameters. The advantage of this approach is that it does not require iterations nor the selection of tuning parameter(s). A potential drawback of this approach though, particularly when the number of observations is small, is that the quality of the final estimation is determined by the qualities and convergence rates of the nonparametric estimates. Consequently, they did not fully utilize the ODE structure in variation reduction.

One main concern of RHCC is the labor-intensive step of tuning parameter selection in the penalized optimization component. After some theoretical and numerical investigations, we found that this tuning parameter selection step is not only computationally burdensome, it can also inflate estimation variation. We thus propose to eliminate this step. Since we do not need to select the tuning parameter, our method is fast, stable and free of some potential drawbacks due to an imperfectly selected tuning parameter. Consequently, it has a great potential to be extended to more complex setup when further regularization is necessary. In our numerical simulation studies, we used cubic B-splines as the basis functions in all approaches. We found that, in comparison to RHCC, the simplified approach is much less sensitive to the number of knots used in the B-spline fitting. This phenomenon is the opposite to that of penalized splines in which the penalty term reduces the sensitivities toward the number of knots used.

These intriguing numerical findings lead us to investigate the structure of the inner step on the quality of final estimates. The rest of this chapter is organized as follows. In Section 3.2, we outline the study framework and conduct analytical investigations on relationships between estimation variation and the structure of inner step of the RHCC procedure. Based on our investigations, we propose a specific version of RHCC procedure with modifications. The consistency and the asymptotic normality properties of the proposed estimators are provided in Section 3.3. In Section 3.4, we compare our method with the original RHCC by simulation studies. We first study the FitzHugh-Nagumo model, and then investigate the compartment models, which are frequently used in modeling the dynamics in ecosystems. A predator-prey dynamic model for lynx and hare is studied in Section 3.5. Finally we conclude this chapter with some remarks in Section 3.6.

3.2 Estimation Method and Variance Comparisons

We consider the following setup, which is as given in RHCC with some slight modifications. Let $\phi_i(t)$ denote the basis functions for the i-th component curve, $X_i(t)$, of the ODE dynamic system and assume that $X_i(t)$ can be approximated by $\hat{X}_i(t)$ through a basis function expansion:

$$\hat{X}_i(t) = \sum_{k=1}^{K_i} c_{ik}\phi_{ik}(t) = \mathbf{c}_i^{\mathrm{T}}\phi_i(t). \tag{3.1}$$

Typically, one chooses B-spline basis or Fourier basis functions. The ODE curves are controlled by ODE parameters and initial values. Often, the initial values play an important role. Differing from the regular setup in RHCC, we include the initial values as part of the parameters and estimate them simultaneously with the rest of the ODE parameters. That is, we let $\theta^* = \{\theta^{\mathrm{T}}, X^{\mathrm{T}}(0)\}^{\mathrm{T}}$.

The procedure includes two steps:

- Inner Step: Given λ and θ^*, and being subject to $\hat{X}(0) = X(0)$, the inner step finds

$$\begin{aligned}
\hat{\mathbf{c}}(\theta^*, \lambda) &= \arg\max_{\mathbf{c}} \sum_{i=1}^{m} J_i(c, \lambda); \\
J_i(c, \lambda) &= \int w_i \ell_i(Y_i - \widehat{X}_i) - \mathrm{PEN}_i(\widehat{X}_i|\lambda_i); \\
\mathrm{PEN}_i(\widehat{X}_i|\lambda_i) &= \lambda_i \int \left\{ d\hat{X}_i/dt - F_i(\hat{X}, \theta) \right\}^2 dt,
\end{aligned} \tag{3.2}$$

  where $\hat{X}_i(t)$ is defined in (3.1); $\ell_i(\cdot)$ is the log-likelihood function of ϵ_i, and w_i's are adjusted normalizing weights with the purpose of making the numerical magnitudes of different components comparable. When $X(0)$ is given and B-splines basis functions are chosen to expand $X(t)$, we carry out the constraint of $\hat{X}(0) = X(0)$ by simply fixing the first element in $\hat{\mathbf{c}}_i$, $\hat{\mathbf{c}}_{i,1}$ to fulfill the constrain.

- Outer Step: The outer step carried out a profile estimation by finding

$$\hat{\theta}^* = \arg\max_{\theta^*} \sum_{i=1}^{m} w_i \ell_i [Y_i - \widehat{X}_i\{t; \hat{\mathbf{c}}(\theta^*, \lambda)\}]$$

  where $\hat{\mathbf{c}}(\theta^*, \lambda)$ is obtained in the inner step. If the observation errors are normally distributed, we can replace ℓ_i with the $(-1) \times L_2$-norm square.

The tunning parameter, λ_i, controls a trade-off between goodness-of-fit to the data and the utilization of the ODE structure. Ramsay et. al (2007) stated that the first term in (3.2), namely the log-likelihood term in the inner step, would help the estimations when the ODE system is mis-specified. Our numerical experiences in §4 leads us to evaluate the problem from an alternative view. When the number of Y is small, the dominating variation in estimating $X(t)$ and θ could result from this term of fitting to Y in (3.2). When an additional penalty term is needed for various regularization purposes, this added variation in the inner step frequently leads to less stable procedures. We focus on having a scaler Y with normally distributed errors and let $\alpha = \lambda^{-1}$, so that $\alpha = 0$ corresponds to the case that the inner step (3.2) contains no elements from Y. Below, for all values of λ, we use two simple models, linear and compartmental models, to illustrate how the variances in RHCC estimator decrease with α. When α belongs to a neighborhood of 0, we are able to show that, asymptotically, the smallest variation of $\widehat{\theta}^*$ is achieved when $\alpha = 0$. Adopting a large λ, or equivalently a small α, is recommended in all publications using the RHCC procedure. Consequently, this is a scenario that is of interest the most. Additional numerical evidences for finite samples are provided in §3.4 through simulation studies.
<u>Linear Model:</u> We consider the ODE equation $dX(t)/dt = a$ and let Y be observed with normal errors with mean 0 and standard deviation σ. We show that the variance of the estimator $\hat{a}$ satifies the following proposition.

Proposition III.1. *Let $\hat{a}$ denote the estimator given in §2 with the tuning parameter*

$\lambda = \alpha^{-1}$ *and the inner steps gives*

$$\hat{X}_a = \arg\min_{X_a} \alpha \|Y - X_a(\mathbf{t})\|^2 + \int \left(\frac{dX_a}{dt} - a \right)^2 dt.$$

Then variance of $\hat{a}$ is a non-decreasing function of α.

A sketch proof of this proposition is given in the Appendix B. From the above proposition, we know that the variance of the estimator with (3.2) as the inner step is minimized when $\alpha = 0$. That is, the inner step does not contains the term involving Y.

Compartmental Model: The simple one-component model fulfills the ODE equation $dX(t)/dt = aX(t)$. Let $\hat{a}$ denote the estimator with the inner step being,

$$\hat{X}_a = \arg\min_{X_a} \alpha \|Y - X_a(\mathbf{t})\|^2 + \int \left(\frac{dX_a}{dt} - aX \right)^2 dt.$$

The close-form expression for the exact variance of $\hat{a}$ cannot be easily derived. In the Appendix B, for any given α, we derived an approximation of the variance of $\hat{a}$. We then evaluate the relationship between α and this variance; Figure 3.1 illustrates these relationships for three values of a. We again observe that the variance of $\hat{a}$ is a non-decreasing function of α. In Section 4, we extend the variance investigation by simulating from a two-compartment model and allow the assumed model to be either correctly or incorrectly specified.

General Model, Small α: The variance properties described above hold for any given $\alpha \geq 0$. The problem of most interest is when λ is large or equivalently when α is close to zero. Under this scenario, we can obtain a general approximation for asymptotic variance of $\widehat{\theta}^*$ when the inner step of the procedure is given by

$$\hat{X}_a = \arg\min_{X_a} \alpha \|Y - X_a(\mathbf{t})\|^2 + \int \left\{ \frac{d\Phi}{dt} c - F(\Phi c, \theta^*) \right\}^2 dt.$$

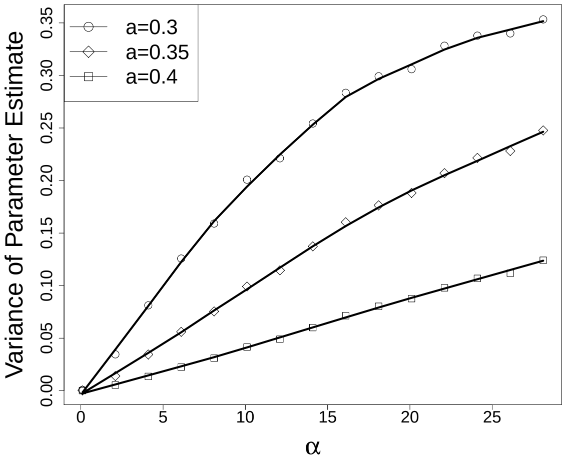

Figure 3.1: The relationship between the variance of estimated parameter, $\hat{a}$, and the values of α at different choices of a; $a = 0.3$, 0.35 and 0.4, respectively.

We show that the variance of the estimator $\hat{\theta}^*$ satifies the following proposition.

Proposition III.2. *When α is close to zero and $\widehat{\theta^*} - \theta^*$ goes to zero as n goes to ∞, the asymptotic variance of $\widehat{\theta^*}$ is minimized when $\alpha = 0$.*

A sketch proof of this proposition is given in the Appendix B. What we did was to put the asymptotic expansion into a particular structure so that we could borrow strength from the semiparametric efficiency theory to show that any additional terms corresponding to a non-zero α would cause extra variation. In particular, by not having any response Y from the log-likelihood term in the inner step, the RHCC procedure with an infinite λ actually reaches the most efficient solution for the given model. On the other hand, when α is non-zero, through it, a set of additional terms that involve the derivatives of spline coefficients, c, with respect to θ^* and α, or the derivatives of L2 distance between dX/dt and the assumed ODE structure with respect to spline coefficients, c, could all contribute to the inflated variation. This theoretical observation is reflected by the additional variation that is associated with choices of the number of basis functions in spline fitting for non-zero α. This phenomenon differs from what is consistently observed in penalized spline literature, in which the additional penalty term reduces the sensitivity to the number of basis functions. We illustrate this numerically in §3.4.

3.3 Asymptotic Properties

Hereafter, we refer to the modified RHCC approach, which does not contain the log-likelihood term in the inner step but includes the initial values as part of the parameters, as the proposed approach. It is an approach that has potential to further incorporate additional regularization term. In this section, we denote $\theta^* = \{\theta^{\mathrm{T}}, X^{\mathrm{T}}(0)\}^{\mathrm{T}} (\in (\Theta \times \Gamma))$ and report asymptotical properties of the estimator $\hat{\theta}^*_n$. We show that the proposed estimator is consistent and asymptotically normal. To ease the presentation, we assume that all components in $X(t)$ are observed at the same time points, T_j, $j = 1, \cdots, n$; T_j are independent random variables on $[0, T]$ that have a distribution Q and density q. We also assume that random vectors, $(\varepsilon_{11}, \cdots, \varepsilon_{m_o 1}, T_1)$, $\cdots$, $(\varepsilon_{1n}, \cdots, \varepsilon_{m_o n}, T_n)$ are independent, identically distributed, where $\varepsilon_{ij} = \varepsilon_i(T_j)$. Further, we denote, by $\tau^{(n)}$, the knots $(0 = t_1^{(n)} \leq \cdots t_{k_n}^{(n)} = T)$, and, by B_n, the space spanned by cubic spline functions corresponding to $\tau^{(n)}$. Precisely, for a given θ^*, we find, in the inner step, $\hat{X}_{ni}(\theta^*, t) = v_i(t)$ among $v_i(t) \in B_n$ such that $v_i(0) = X_i(0)$ and $\hat{X}_n(\theta^*, t)$ minimizes the objective function in (3.2). We then obtain $\hat{\theta}^*_n$ in the outer step. Qi and Zhao (2010) reported asymptotic properties of the RHCC estimators with non-zero α under certain conditions. Our Lemmas and Theorems follow the structure of those in Qi and Zhao (2010), and the proofs can be carried out analogously, taking into account the differences in the inner steps. The proofs are thus omitted, but they can be obtained from the first author.

Under the main assumptions given in the Appendix B, we have the following results.

Lemma III.3. *Under Assumption B2, for any compact set Θ_0 of Θ and any compact* *subset* Γ_0 *of* Γ*, we have, for all* i*,*

$$\lim_{n\to\infty} \sup_{\theta^*\in\Theta_0\times\Gamma_0} \inf_{v_i\in B_n, v_i(0)=X_i(0)} \|X_i(\theta^*,t)-v_i(t)\|_\infty \vee \left\|\frac{dX_i}{dt}(\theta^*,t)-\frac{dv_i}{dt}(t)\right\|_\infty \;=\; 0,$$

where $a\vee b$ *means* $\max(a,b)$ *for any real numbers* a *and* b*.*

Further, denote

$$r_n \;=\; \max_i \left\{ \sup_{\theta^*\in\Theta_0\times\Gamma_0} \inf_{v_i\in B_n, v_i(0)=X_i(0)} \|X_i(\theta^*,t)-v_i(t)\|_\infty \vee \left\|\frac{dX_i}{dt}(\theta^*,t)-\frac{dv_i}{dt}(t)\right\|_\infty \right\}.$$

Lemma III.4. *Under Assumptions B.1-B.4, suppose* $\sum_{i=1}^m w_i = m$*, then for any compact subset* Θ_0 *of* Θ *and any compact subset* Γ_0 *of* Γ*, we have, for all* i *and large enough* n*,*

$$\sup_{\theta^*\in\Theta_0\times\Gamma_0} w_i\|\hat{X}_{ni}(\theta^*,t)-X_i(\theta^*,t)\|_\infty \;\le\; mT\sqrt{4m(8K^2+2)}r_n e^{mKT}, \tag{3.3}$$

where K *is a constant depending only on the set* $\Theta_0\times\Gamma_0$ *and the function* F*.*

Theorem III.5. *Suppose that Assumptions B1, B2, B3 and B5 hold and that* $\hat{\theta}^*_n$ *is uniformly tight. Denote the true parameter vector as* θ^*_0*. Then the estimator* $\hat{\theta}^*_n$ *is consistent, i.e.,* $\hat{\theta}^*_n \to \theta^*_0$ *in probability.*

Theorem III.6. *Suppose that Assumptions B1, B2, B3 and B6 hold and that* $\hat{\theta}^*_n$ *is uniformly tight. Let* $\ell(Y(t),X(\theta^*,t)) = \sum_{i=1}^m w_i\ell_i(Y_i(t),X_i(\theta^*,t))$ *be the weighted log-likelihood function of the observed error at* $(Y(t),X(\theta^*,t))$ *with weight* w_i *for the i-th component* $(Y_i(t),X_i(\theta^*,t))$*. Suppose that* $r_n = o_p(1/n)$ *and* $V_{\theta^*_0}$ *is non-singular, where*

$$\begin{aligned} V_{\theta^*_0} \;&=\; E_{\theta^*_0}\left\{\sum_{i=1}^m w_i\left[\frac{\partial\ell_i}{\partial X_i}\{Y_i(t),X_i(\theta^*_0,t)\}\frac{\partial^2 X_i}{\partial\theta^*\partial\theta^{*T}}(\theta^*_0,t)\right.\right. \\ &\quad \left.\left. + \frac{\partial^2\ell_i}{\partial^2 X_i}\{Y_i(t),X_i(\theta^*_0,t)\}\frac{\partial X_i}{\partial\theta^*}(\theta^*_0,t)\frac{\partial X_i}{\partial\theta^{*T}}(\theta^*_0,t)\right]\right\}. \end{aligned} \tag{3.4}$$

Then $\sqrt{n}(\hat{\theta}^* - \theta_0^*)$ *is asymptotically normal with mean zero and the asymptotic covariance matrix is given by*

$$V_{\theta_0^*}^{-1} E_{\theta_0^*} \left\{ \sum_{i=1}^{m} w_i \left(\left[\frac{\partial \ell_i}{\partial X_i} \{Y_i(t), X_i(\theta_0^*, t)\} \right]^2 \left[\frac{\partial X_i}{\partial \theta^*}(\theta_0^*, t) \right]^{\mathrm{T}} \left[\frac{\partial X_i}{\partial \theta^*}(\theta_0^*, t) \right] \right) \right\} V_{\theta_0^*}^{-1}.$$

The definition of uniform tightness is given as in Qi and Zhou (2010) and is restated in the Appendix. The consistency and normality results mainly rely on Lemma III.4, which shows the approximate rate for the B-spline approximation to the true ODE curves for any given θ^*. In the RHCC with a non-zero $\alpha = \lambda_n^{-1}$, the rate for the equivalent step is

$$\left[O_p\left(\frac{1}{\sqrt{\lambda_n}} \right) \sqrt{T} + mT\sqrt{4m(8K^2+2)r_n} \right] e^{mKT}. \tag{3.5}$$

When the log-likelihood term exists in the inner step, the convergence rate is determined by the balance betweem $1/\sqrt{\lambda_n}$ and the rate of r_n. In numerical studies, we can control or reduce the r_n rate by choosing a larger number of knots. However, as long as one needs to choose λ_n in the inner step, as having been emphasized in the literature, this task needs to be carried out carefully. For some choices of λ_n, the total rate in (3.5) might not be improved by using more knots. The proposed method will have less of this concern since the rate (4.9) is only determined by r_n.

3.4 Simulation Study

In this section, we apply the proposed method to two simulated ODE dynamic systems and compare the results with those of RHCC with different choices of λ_n. Both ODE systems in the simulations have been used in ample of scientific applications.

EXAMPLE 1. In the first simulation, we consider the FitzHugh-Nagumo ODE model (FitzHugh, 1961; and Nagumo et al., 1962). This model was developed to simplify the Hodgkin-Huxley model (1952) for the behavior of spike potential in the giant axon of squid neurons. The ODEs are

$$\begin{aligned} \frac{dV}{dt} &= c\left(V - \frac{V^3}{3} + R\right), \\ \frac{dR}{dt} &= -\frac{1}{c}(V - a + bR), \end{aligned} \tag{3.6}$$

where V describes the voltage across an axon membrane, R is the recovery variable summarizing outward currents, and a, b, c are parameters in the dynamic system.

We let $a = 0.2$, $b = 0.2$ and $c = 3.0$, respectively, and generate V and R from (4.12) with initial conditions $V(0) = -1.0$ and $R(0) = 1.0$. We let errors for both V and R be from the normal distribution with standard deviation $\sigma = 2$ and simulate 201 pairs of them in the time interval $(t \in)[0, 20]$. We place 201 knots in the range of $[0, 20]$ with equal distance between two consecutive knots. The cubic B-splines are used to approximate the ODE curves.

We consider two ways to calculate the estimated values for $X_i(t)$ $(i = 1, \cdots, m)$, which are $V(t)$ and $R(t)$ in this example. First, we use the B-spline estimates, $\hat{X}_i$, in the inner step corresponding to the final estimate of θ^*, $\hat{\theta}^*$. As an alternative, we use the ODE solution to equations (4.12) with $\theta^* = \hat{\theta}^*$. We denote these solutions by $X_{i,\hat{\theta}^*}$, $i = 1, 2$. In general, for an m-component ODE system, we define

$$\begin{aligned} \text{MSE}_1 &= \frac{1}{mn}\sum_{i=1}^{m} w_i \|\hat{X}_i(\mathbf{t}_i) - X_i(\mathbf{t}_i)\|^2, \\ \text{MSE}_2 &= \frac{1}{mn}\sum_{i=1}^{m} w_i \|X_{i,\hat{\theta}^*}(\mathbf{t}_i) - X_i(\mathbf{t}_i)\|^2. \end{aligned} \tag{3.7}$$

We note that the lower the value of MSE_1 is, the more precise are the estimated values of ODE components, while the lower the value of MSE_2 is, the more precise are the estimated values of the ODE parameters and initial values. We also compare the mean square errors for the estimated first derivatives of the ODE components.

They are analogously defined as

$$\begin{aligned}
\text{DMSE}_1 &= \frac{1}{mn}\sum_{i=1}^{m} w_i \|\hat{X}_i'(\mathbf{t}_i) - X_i'(\mathbf{t}_i)\|^2, \\
\text{DMSE}_2 &= \frac{1}{mn}\sum_{i=1}^{m} w_i \|X_{i,\hat{\theta}^*}'(\mathbf{t}_i) - X_i'(\mathbf{t}_i)\|^2. \qquad (3.8)
\end{aligned}$$

We also compare the estimated ODE parameters and initial values produced by the proposed and RHCC methods, respectively. For the RHCC method, the following tuning parameter selection method was proposed in Ramsay et al. (2007):

$$\lambda_{selected1} = \arg\min_{\lambda} \sum_{i=1}^{m} w_i \|\hat{X}_i(\mathbf{t}_i) - X_{i,\hat{\theta}^*}(\mathbf{t}_i)\|^2. \qquad (3.9)$$

It is a common practice to use this selection criterion to find a sufficiently large but not too large λ that obtains a local minimum. We further consider the following alternative that minimizes the weighted squared prediction error (PE) defined as:

$$\lambda_{selected2} = \arg\min_{\lambda} \sum_{i=1}^{m} w_i \|Y_i(\mathbf{t}_i) - X_{i,\hat{\theta}^*}(\mathbf{t}_i)\|^2. \qquad (3.10)$$

Both crieria were used throughout the numerical investigations for RHCC.

We repeat the simulation 100 times and report in Table 3.1 the results for the estimated ODE parameters and initial values, MSEs for the estimated $V(t)$ and $R(t)$ and DMSEs for the estimated derivatives. Based on the above tuning parameter selection criterion, the average logarithm of the tuning parameter in RHCC method is around 6.4 through the tuning parameter selection (3.9), and around 4.1 through (3.10). In this example, using a λ much larger than $10^{4.1}$ increases the estimation variation. As we can see (from the bottom half of Table 3.1), in terms of MSE and DMSE, the proposed method performs better than RHCC. Within the RHCC results under two different tuning parameter selections, the selection criterion (3.9) tends to have smaller MSE and DMSE results. In terms of parameters and initial values estimation, the outcomes for RHCC with the best selected λ and those for the proposed method are roughly comparable, though the variances of the estimated ODE parameters and initial values obtained by the proposed method are smaller than the ones obtained by the RHCC method.

We further investigate this ODE system in the first mis-specified scenario. We assume that the ODE parameters a, b and c are perturbed shortly during the time range $[0, 20]$ but remain constants for the rest of the range. Precisely, the parameters are,

$$\begin{aligned}
a &= 0.2I(0 \le t \le 3) + \frac{0.2}{9}[18 - (t-6)^2]I(3 < t \le 9) + 0.2I(9 < t \le 20), \\
b &= 0.2I(0 \le t \le 11) + \frac{0.2}{9}[18 - (t-14)^2]I(11 < t \le 17) + 0.2I(17 < t \le 20), \\
c &= 3.0I(0 \le t \le 7) + \frac{1.5}{9}[9 + (t-10)^2]I(10 < t \le 13) + 3.0I(13 < t \le 20),
\end{aligned}$$

where I is the indicator function. We let initial values, observed times, observation errors and choices of knots to be the same as those in the above correctly specified setting. The outcomes over 100 simulations are reported in Table 3.2. The true ODE parameters and the average estimates using the proposed and the RHCC methods are given in Figure 3.2. We note that the MSE results are similar for the proposed and RHCC methods and the estimation qualities of the ODE parameters are also comparable.

	Estimated Parameters				
	a	b	c	$V(0)$	$R(0)$
True	0.2	0.2	3.0	-1.0	1.0
Proposed	0.201(0.090)	0.182(0.305)	2.879(0.167)	-0.936(0.336)	1.002(0.299)
RHCC[1]	0.193(0.083)	0.225(0.339)	2.843(0.196)	-0.924(0.369)	0.898(0.343)
RHCC[2]	0.202(0.094)	0.177(0.314)	2.875(0.170)	-0.890(0.366)	0.962(0.350)

	MSE and DMSE			
	MSE_1	MSE_2	DMSE_1	DMSE_2
Proposed	0.0096	0.0096	0.0558	0.0558
RHCC[1]	0.0138	0.0139	0.0700	0.0701
RHCC[2]	0.0123	0.0110	0.0612	0.0602

Table 3.1: Simulation results for Example 1. [1]The tuning parameter is selected using the criterion (3.9) ; [2]The tuning parameter is selected using the criterion (3.10). Top half of the table: Monte Carlo means and Monte Carlo standard deviations (inside the parentheses) for estimated ODE parameters and initial values for the proposed and RHCC methods. Bottom half of the table: MSE and DMSE values, calculated by (3.7) and (3.8), for the proposed and RHCC methods.

	Estimated Parameters				
	a	b	c	$V(0)$	$R(0)$
True	-	-	-	-1.0	1.0
Proposed	0.162(0.097)	0.264(0.255)	2.305(0.148)	-1.678(0.628)	1.422(0.499)
RHCC[1]	0.154(0.095)	0.295(0.281)	2.282(0.162)	-1.419(0.732)	1.204(0.612)
RHCC[2]	0.158(0.098)	0.271(0.258)	2.298(0.158)	-1.486(0.653)	1.336(0.570)

	MSE and DMSE			
	MSE_1	MSE_2	DMSE_1	DMSE_2
Proposed	0.0227	0.0227	0.0893	0.0899
RHCC[1]	0.0262	0.0262	0.0944	0.0945
RHCC[2]	0.0236	0.0242	0.0881	0.0920

Table 3.2: Simulation results for Example 1 with an incorrectly specified model. Entries are as in Table 1.

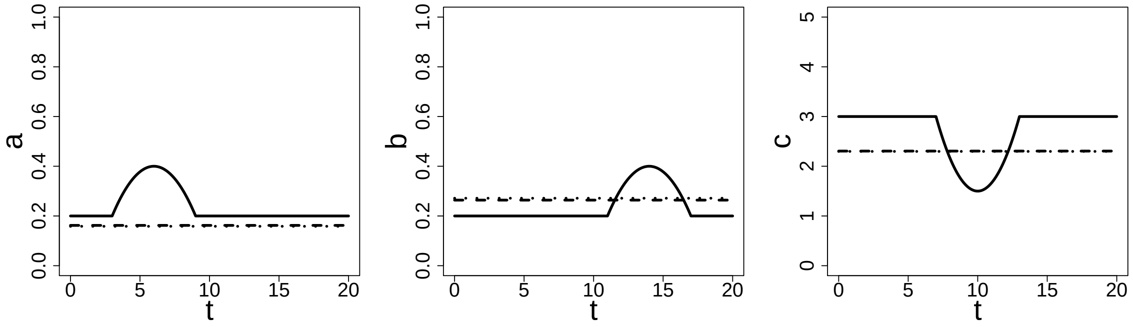

Figure 3.2: The ODE parameters and their estimates for Example 1 when model is incorrectly specified. Solid lines: the true non-constant ODE parameters; Dashed lines: the estimates by the proposed method; Dotted lines: the estimates by the RHCC method with the tuning parameter selected using (3.10).

EXAMPLE 2. In the second example, we use dynamic compartment models developed to model dynamic behavior of pollutants in ecosystems (Bulter, 1978; Neely, 1980; Dickson et al., 1982) or to describe biogeochemical cycles in an ecosystem (Hutzinger, 1985). The simplest possible compartment model is obtained by assuming the function vector F in Section 3.1 is linear, and the general ODEs for this compartment model are as follows,

$$\frac{dX_i(t)}{dt} = \sum_{i'=1}^{m} k_{i'i} X_{i'}(t) + F_{0i}, \text{ for } i \ = \ 1, \cdots, \ m,$$

where $k_{i'i}$ and F_{0i} are parameters in this model. We write the above equations in a matrix notation as, $dX/dt \ = \ AX + B$, where A and B are parameter matrix functions of $k_{i'i}$ and F_{0i}. Since the function vector F is linear, the system has an analytic solution and the parameter estimation can be carried out by the least square method without relying on the algorithms discussed in this paper. We consider it here for illustrative purposes and we can also easily study the performances of the procedures when the model is mis-specified. In this simulation, we will consider a two-compartment dynamic model with the corresponding ODEs being

$$\frac{dX_1}{dt} = k_{11}X_1 + k_{12}X_2, \text{ and } \frac{dX_2}{dt} = k_{22}X_2, \tag{3.11}$$

where k_{11}, k_{12} and k_{22} are parameters. The parameter k_{12} describes the interaction between the two compartments of the system. If it is zero, there is no interaction inside the system, and the true ODE can be written as

$$\frac{dX_1}{dt} = k_{11}X_1, \text{ and } \frac{dX_2}{dt} = k_{22}X_2. \tag{3.12}$$

First we consider the scenario that the ODEs are correctly specified. We generated data from the ODE (3.11) with the true parameter values $k_{11} = 0.3$, $k_{12} = 0.2$ and

$k_{22} = 0.2$. We set both initial values, $X_1(0)$ and $X_2(0)$, to be one. The observation errors of X_1 and X_2 are added as: $Y_{X_\ell,j} = X_\ell(t_j) + \varepsilon_{\ell j}$, $\ell = 1,\ 2$, where ε_{1j} and ε_{2j} are independent standard normal random variables with mean 0 and standard deviation one. We simulated 201 pairs of observations from the two compartments in the time interval [0,10] on equally-distanced grids. The ODE curves were estimated by the cubic B-splines with knots at each time point t_j. We used the reciprocals of the variance taken over the simulated observed data as the weight w_i for the i-th component. The simulations were performed over 500 repetitions. Besides comparing the estimated ODE parameters and initial values for different estimation approaches, we also compare the mean square errors (MSE) for the estimated values of each ODE component.

For RHCC, the best performance on Example 2 is observed at the upper limit of the tuning parameter value allowed by the software ($\lambda = 10^{12}$). This choice is suggested by both (3.9) and (3.10). In comparisons to the outcomes of RHCC with this tuning parameter, we found that our method performs similarly to RHCC; the relative differences in means of estimated parameters and the associated Monte Carlo standard deviations (MCSD) are 4% or less (not shown) with the proposed method having slightly smaller MCSD's. From (4.9) and (3.5), if λ_n is very large, the two rates are similar and dominated by the rate of r_n, which is controlled by the knots, which explains the similarities. The MSE results for the proposed method are also slightly better than those for the RHCC method, with the MSE and DMSE for the two methods being ($\times 10^3$) 1.217 versus 1.222 and 1.946 versus 2.179 respectively. The two formulae in (3.7) and (3.8) produce identical values for MSE and DMSE for both methods.

To investigate the performance of the RHCC method with other values of the tuning parameter, we artificially set λ to be numbers smaller than 10^{12}, and produce boxplots for the estimated ODE parameters and initial values over 500 repetitions.

The results are given in Figure 3.3. As expected, the variance of the estimates is larger for a smaller λ, and so are the values of MSE (not shown).

We then study how the estimation results are affected by the number of knots in this simulation. Recall that the number of knots is denoted by k_n. Let the distance between two consecutive knots be $10/(k_n - 1)$ for the whole rage of [0, 10], and let $(k_n - 1)$ vary from 20 to 80. Other settings are maintained exactly the same as before. We plot the means of the estimated ODE parameters and initial values in Figure 3.4. It can be seen that for moderate values of k_n, the proposed method is much less sensitive to the number of knots than the RHCC method.

To further investigate the performance of our method and the RHCC method in a mis-specified ODE system, we simulate the data from the compartment model with a non-zero interaction term, k_{12}, in (3.11) but perform the parameter estimation assuming (3.12). In (3.11), we fix $k_{11} = 0.3$ and $k_{22} = 0.2$ throughout. We then let k_{12} vary among 0.1, 0.3 and 0.5, respectively, to generate 201 pairs of X_1 and X_2 with initial values $X_1(0) = 1$ and $X_2(0) = 1$. We maintain other settings the same as before. For RHCC, the best λ remains to be 10^{12}. The Monte Carlo means and standard deviations for the estimated parameters and the two pairs of MSE and DMSE values are reported in Table 3. The proposed and RHCC methods perform similarly, with the Monte Carlo standard deviations, MSEs and DMSEs for the RHCC being slightly larger than those for the proposed method when $k_{12} = 0.1$ and 0.3, and being slightly smaller when $k_{12} = 0.5$.

We observe, in this example, that regardless the assumed model being correctly specified or not, the proposed method either performs similarly to RHCC with the "best" choice of λ or outperforms it for other choices of λ. The asymptotic results seem to hold well in finite sample scenarios. Furthermore, the proposed method eliminates additional computational efforts to select λ.

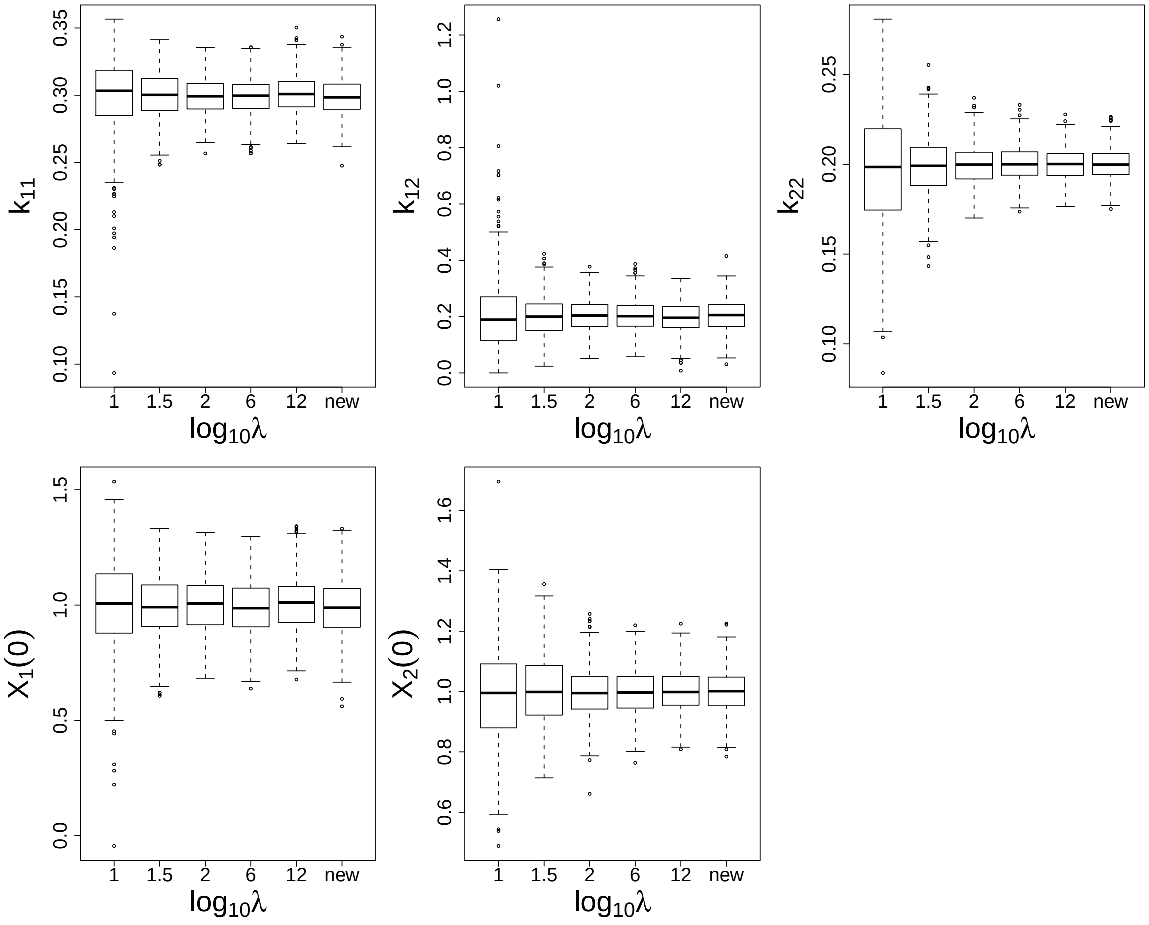

Figure 3.3: Boxplots for the estimated ODE parameters and initial values. For the RHCC method, different values of the tuning parameter λ are used as the label in the X-axis; while the proposed method is indexed by "new".

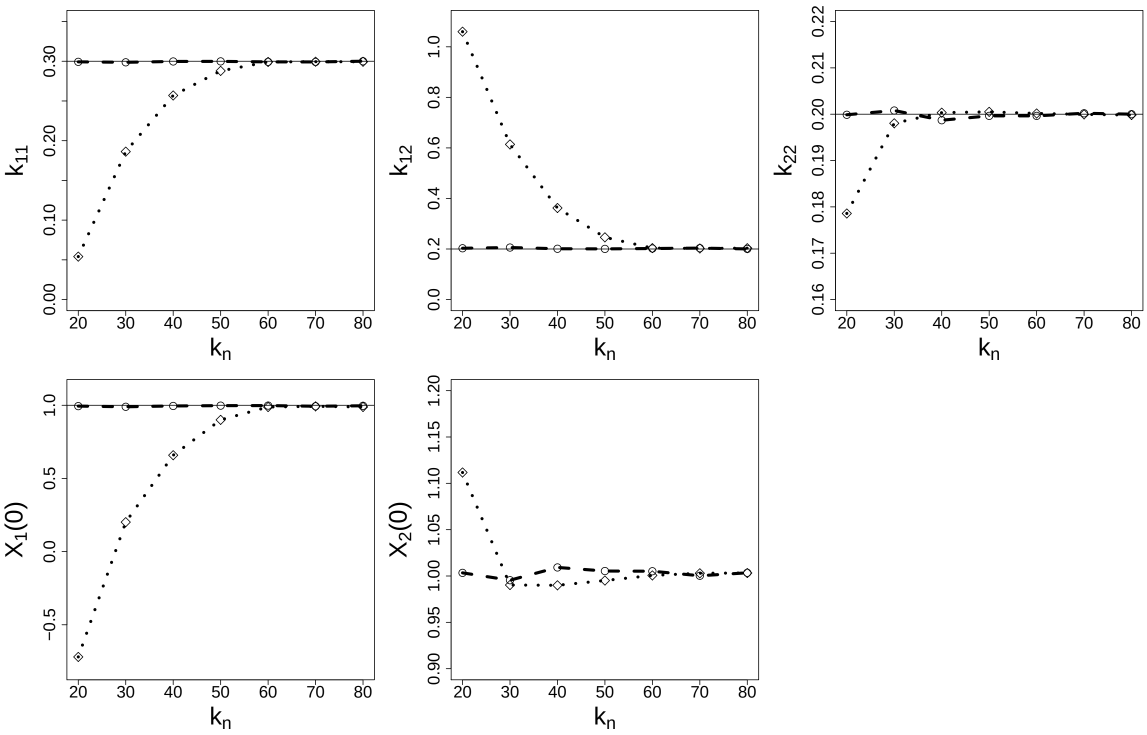

Figure 3.4: The means of the estimated ODE parameters and initial values produced by the proposed method (dashed lines and circle) and by RHCC (dotted lines and diamond) with different numbers of k_n, where k_n = the number of knots $-$ 1. The solid horizontal line indicates the true parameter value.

			Estimated Parameters		
		k_{11}	k_{22}	$x_1(0)$	$x_2(0)$
	True	0.3	0.2	1.0	1.0
$k_{12} = 0.1$	Proposed	0.3333(0.0037)	0.1998(0.0092)	1.1797(0.0381)	1.0030(0.0722)
	RHCC	0.3332(0.0039)	0.2000(0.0096)	1.1816(0.0404)	1.0030(0.0777)
$k_{12} = 0.3$	Proposed	0.3588(0.0022)	0.2004(0.0089)	1.6325(0.0314)	0.9982(0.0711)
	RHCC	0.3592(0.0022)	0.2001(0.0093)	1.6286(0.0317)	1.0012(0.0751)
$k_{12} = 0.5$	Proposed	0.3700(0.0016)	0.2005(0.0097)	2.1061(0.0302)	0.9984(0.0773)
	RHCC	0.3698(0.0016)	0.2006(0.0094)	2.1093(0.0300)	0.9954(0.0741)

			MSE and DMSE		
		MSE_1	MSE_2	DMSE_1	DMSE_2
		$(\times 10^2)$	$(\times 10^2)$	$(\times 10^4)$	$(\times 10^4)$
$k_{12} = 0.1$	Proposed	0.13	0.13	2.885	2.885
	RHCC	0.14	0.14	3.011	3.011
$k_{12} = 0.3$	Proposed	0.15	0.15	4.728	4.728
	RHCC	0.16	0.16	4.976	4.976
$k_{12} = 0.5$	Proposed	0.18	0.18	6.602	6.602
	RHCC	0.17	0.17	6.364	6.364

Table 3.3: Simulation results for Example 2 with an incorrectly specified model. Entries are as in Table 1.

3.5 Analysis of Lynx and Hare Data

The numbers of trapped lynx and snowshoe hares from North Canada between 1900 and 1920 were collected by the Hudson Bay company (Odum 1953); the observed data should reflect their relative populations. In this section, we apply the proposed and RHCC methods and fit this data set with the Lotka-Volterra dynamic model (Lotka 1910 and Volterra 1926), which is the most commonly used predator-prey model for two species. The model has two components, described by the following ODEs:

$$\frac{dH}{dt} = aH - bHL, \quad \frac{dL}{dt} = -cL + dHL,$$

where a, b, c and d are parameters. For the lynx-hare data set, H represents the evolution function of the number of snowshoe hares and L for the number of lynx. The starting values (needed for numerical optimization) of $H(0)$, $L(0)$ and the ODE parameters a, b, c and d were obtained by replacing H, L and their derivatives by the corresponding spline estimates and then using the given structures to solve for them. We use cubic spline basis with 201 knots for both methods.

For RHCC, unlike in simulation studies, criteria (3.9) and (3.10) suggested very different values for λ. The criterion (3.10) suggested $\lambda = 10^5$ for this data set. With this λ, the RHCC method practically obtained almost identical results as those of the proposed method. On the other hand, the criterion (3.9) suggested $\lambda = 10^{12}$. This λ led to estimated curves that deviate from the data points, particularly in the areas that are near the peaks. The estimated H and L curves by RHCC with $\lambda = 10^5$ and 10^{12}, and by the proposed method are provided in Figure 3.5. The outcomes for the estimated ODE parameters and initial values are summarized in Table 3.4. A bootstrap analysis by re-sampling residuals 100 times were used to calculate estimated standard errors, and they are reported inside the parentheses besides the corresponding estimates. Table 4 also shows how the results of RHCC vary with λ. The proposed method and RHCC with $\lambda = 10^5$, selected using the criterion (3.10), give the best result.

To study the effects due to the number of knots, in Figure 3.6, we plot the values of weighted PE against $k_n - 1$, where k_n is the number of knots. The two curves correspond to the proposed method and RHCC with $\lambda = 10^5$. The PE values for the RHCC method with $\lambda = 10^{12}$ vary between values of 30 to 40 and are not plotted. Similar as being observed in simulation studies, both methods benefit from using a large number of knots and the proposed method is much less sensitive to the number of knots than RHCC.

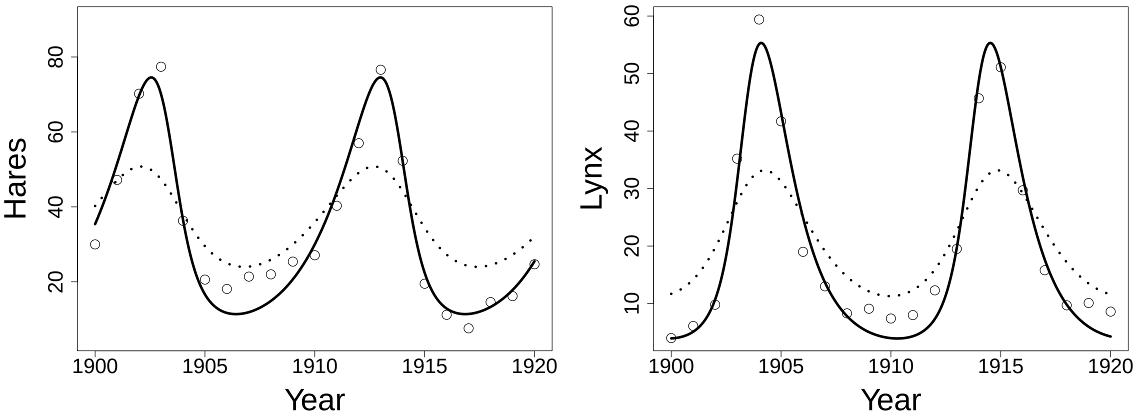

Figure 3.5: Estimated $H(\cdot)$ and $L(\cdot)$ for the hare-lynx data set. Solid lines: the proposed method and the RHCC method with $\lambda = 10^5$; dotted lines: the RHCC method with $\lambda = 10^{12}$.

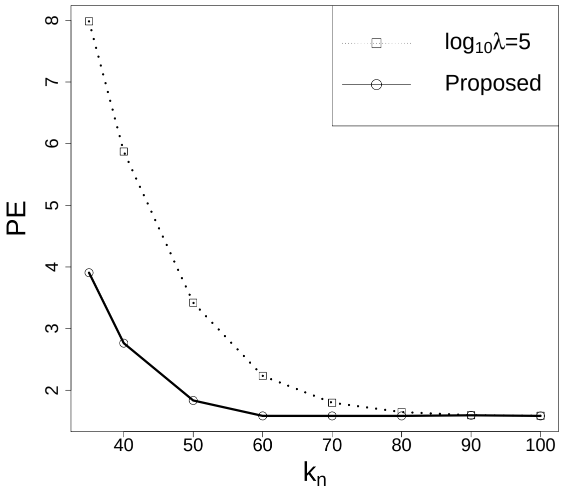

Figure 3.6: The values of PEs versus k_n in the hare-lynx data analysis; k_n: number of knots $-$ 1.

	a	b	c	d	$H(0)$	$L(0)$
Proposed	0.4823	0.0248	0.9177	0.0273	35.4007	3.9377
	(0.0322)	(0.0014)	(0.0593)	(0.0017)	(1.4274)	(0.4434)
RHCC	0.4206	0.0207	0.8630	0.0242	40.2292	11.6887
$\lambda = 10^{12}$	(0.3132)	(0.0166)	(0.6687)	(0.0180)	(7.0825)	(4.4004)
RHCC	0.4831	0.0248	0.9142	0.0272	35.2253	4.1158
$\lambda = 10^{8}$	(0.0352)	(0.0015)	(0.0614)	(0.0017)	(1.6173)	(0.4935)
RHCC	0.4823	0.0248	0.9177	0.0273	35.3990	3.9372
$\lambda = 10^{5}$	(0.0322)	(0.0014)	(0.0593)	(0.0017)	(1.4275)	(0.4434)
RHCC	0.4980	0.0253	0.9163	0.0272	34.5163	3.8239
$\lambda = 10^{2}$	(0.0347)	(0.0014)	(0.0602)	(0.0017)	(1.7254)	(0.5186)
RHCC	0.5664	0.0267	0.9657	0.0273	31.2388	4.5909
$\lambda = 10^{1}$	(0.1615)	(0.0068)	(0.3598)	(0.0079)	(6.4145)	(2.8393)

Table 3.4: The estimated ODE parameters and initial values by the proposed method and by RHCC with different values of λ. The corresponding bootstrap standard errors are given inside the parentheses.

3.6 Concluding Remarks

We investigated the role played by the penalized estimation component in the inner-step of the parameter-cascades estimation method of RHCC in finding ODE solutions. We found that an alternative approach, which still utilizes "collocation" and parameter-cascades estimation but leaves out the part involving the response Y, could result solutions with less variation yet still can be computed without a numerical ODE solver. Conceptually, the inner-step of the proposed approach finds the best approximation under a semiparametric model from all elements in the linear space spanned by cubic B-spline basis functions. When this approximation can be achieved with a very high precision, it is as if one can estimate the nonparametric component in a semiparametric model with a negligible convergence rate, the consequence is equivalent to knowing the nonparametric component. The problem under study then becomes a parametric one and the solutions tend to be less variable and more stable. This is what we have observed in the numerical investigations. Additional advantages of the proposed method include that it is much faster and easier to compute and it is also less sensitive toward the number of knots used to construct the cubic B-spline basis functions. The proposed procedure and the properties we develop here will be useful when other penalty term(s) are imposed for different regularization purposes.

CHAPTER IV

Regularized Semiparametric Estimation for Ordinary Differential Equations

4.1 Introduction

In Engineering, physics and bio-medical science fields, dynamic systems are often modeled through a set of ordinary differential equations (ODEs). Most ODE dynamic systems are fully determined by the parameters and initial values. They usually have nonlinear structures and no trivial analytic solutions. Given the parameters and initial values, there exist various numerical methods to solve nonlinear ODEs, including the well known family of Runge-Kutta methods. In reality, the parameters of the ODE system are often unknown and need to be estimated using the observation data.

Suppose that an ODE dynamic model has the following general structure:

$$\frac{dX}{dt} = F\{X(t), \theta, t\} \tag{4.1}$$

where $X(t) = \{X_1(t), \cdots, X_m(t)\}^T$ is the state vector to describe the dynamic system, $\theta = (\theta_1, \cdots, \theta_d)^T$ denotes the unknown parameters to be estimated, and $F(\cdot) = \{F_1(\cdot), \cdots, F_m(\cdot)\}^T$ is a known force functional structure, which is usually highly non-linear. Instead of directly observing the true state vector $X(t)$, the surrogate

$Y(t)$ is observed at discrete time grids by assuming

$$Y_{ij} = Y_j(t_{ij}) = X_j(t_{ij}) + \varepsilon_{ij},\ i = 1, \ldots, n_j; j = 1, \ldots, m. \tag{4.2}$$

Over a long period of time, the parameters θ were assumed as constants. Currently in the statistics literature, there are two main categories of estimation procedure to solve an ODE system. For a two-stage method, one estimates the ODE curves and their first derivatives in stage-one by a nonparametric smoothing approach, and then, in the second stage, finds the parameter estimates through the classical least-square optimization with $X(t)$ and $dX(t)/dt$ replaced by the nonparametric estimates obtained from the first stage. Varah (1982) estimated $X(t)$ and $dX(t)/dt$ using a spline smoothing technique in stage-one. Recently, Liang and Wu (2008) extended the work of Varah (1982) by using the local polynomial estimation as the smoothing approach and they further provided statistical properties of the estimator. The use of non-parametric kernel estimation method was proposed and studied in Brunel (2008). These approaches are easily implemented and can perform very well with moderate to large data sets with densely observed data points. However, if the level of observation noise is relatively high and/or the sample size is small, the two-stage method may not obtain sufficiently precise estimates of $dX(t)/dt$ in the first stage and consequently the estimation of parameters in the second stage also suffer.

The second category of methods are built on profile estimation. The approach was introduced by Ramsay, et al. (2007), and it has been referred to as a parameter cascade method. Instead of estimating the ODE curves directly from the data, one first constructs the ODE curves as functions of given parameters in the inner step. These estimated functions were then included into the outer step which optimizes an objective function. In Ramsay, et al. (2007) and the follow-up papers, a penalty term is included in the inner step with the intention of balancing the goodness of fit between the observations and the assumed ODE system and the faithfulness toward the assumed system.

Recently, the variation of the approach was investigated in Li, et al. (2011). Their theoretical and numerical findings all suggest that, for variation reduction purpose, one should remove the additional penalty term in the inner step. The resulting simple estimator is the most efficient one for the larger family of estimators considered by Ramsay et al (2007). By considering the ODE initial values as part of parameters and reconstructing the optimization criterion in the inner step, the parameter cascade method gives smaller estimation standard errors and the outcomes are much less affected by certain decision within the nonparametric estimation in the inner step, such as the choice of B-spline knots.

In reality, the parameters θ may not always remain constants as the systems evolute with time. In Chen and Wu (2008), they noticed the ODE parameters in the HIV/AIDS dynamics could vary with time and they applied a two-stage method to estimate the time-varying ODE coefficients. In this paper, we consider a modeling approach that will retain the interpretation advantages of a parametric ODE system, while accommodates potential disturbances that cause the ODE parameters to vary over time. For example, the Lotka-Volterra dynamic model is widely used to study the population evolution of predator and prey in the ecological science. When the two components are dynamically balanced with each other, the parameters of the model are constants. When certain unpredictable human factors or unusual natural phenomena strike, such as earthquake, forest fire or environmentally unsound logging practice, the system may break the balance and the parameter values vary. After a short perturbation due to unusual factors, another balance system may be re-established and the parameters would again be constants. The estimation methods by treating parameters as constants will not be suitable for this situation. Assuming time varying coefficients through out the observation time may result loosing the understandings of the system provided by the constant parameters.

With this setup in mind, we propose a regularized parameter cascade method to estimate the time-varying parameters for ODEs. The main parameter estimation method is based on Chapter III. The variation reduction therein enables a feasible and stable estimation procedure. An equivalent regularization penalty was adopted in James, et al. (2009) to estimate the coefficient function in a linear functional regression model. We assume that there exist time regions that parameters of the ODE system are constants. By adding the penalties in the outer step of a parameter-cascade method, the true parameter structures can be recovered and the estimation variation can be reduced dramatically in comparison to the methods without regularization penalties. We also show that, with probability tending to one and under certain conditions, the distances between estimated ODE curves by the proposed method and the true are bounded at a certain rate as the sample size grows.

The rest of this chapter is organized as follows. In Section 4.2, we propose the estimation method and discuss various important issues in the algorithm including the corresponding degree of freedom and the choice of penalty parameter. The non-asymptotic bounds on the errors of our estimator are presented in Section 4.3. In Section 4.4, we compare our method with other approaches by simulation studies. The two models we investigate are FitzHugh-Nagumo model and Lotka-Volterra model. In Section 4.5, we use the proposed method to analyze a lynx-hare dynamic data set and a measles incidence dynamic data set collected in Ontario Canada. Finally we conclude the paper with a short discussion in Section 4.6.

4.2 Estimation Procedure

In this section, we propose a penalized estimation algorithm and address issues arise in the selection of the penalty tuning parameter.

4.2.1 Algorithm

Letting $\theta_\ell(t), \ell = 1, \dots, d$ denote the time-varing coefficients in an ODE system, and considering a p-dimensional basis $\psi(t) = \{\psi_1(t), \psi_2(t), \dots, \psi_p(t)\}^T$, we consider the following model that allows imperfect fitting:

$$\theta_\ell(t) = \psi(t)^T \eta_\ell + e_{p,\ell}(t) = \xi_\ell(t) + e_{p,\ell}(t),\ \ \ell = 1, \dots, d$$

where η_ℓ is the coefficient vector of the basis expansion, and $e_{p,\ell}(t)$ represents the deviation of $\xi_\ell(t)$ from the true curve $\theta_\ell(t)$. We also denote the space spanned by such a set of basis functions by $\mathbb{L}_{\psi,p}$; $\xi_\ell(\cdot) \in \mathbb{L}_{\psi,p}$. A common choice $\psi(t)$ is the system of B-spline basis Functions, which is what we used. Throughout, we use the notation of $\eta = (\eta_1^T, \dots, \eta_d^T)^T$ and $\theta(\cdot) = \{\theta_1(\cdot), \dots, \theta_d(\cdot)\}^T$.

For the j-th component of ODE curves, we have

$$\hat{X}_j(t) = \phi(t)^T c_j,\ j = 1, \dots, m$$

where $\phi(t) = \{\phi_1(t), \dots, \phi_q(t)\}^T$ is a q-dimensional basis vector and c_j is the coefficient vector. Denote $c = (c_1^T, \dots, c_m^T)^T$. Assuming the observation errors are iid normally distributed and denoting the initial values vector as $X[0] = \{X_1(0), \dots, X_m(0)\}^T$, we let $\eta^* = (X[0]^T, \eta^T)^T$. Following the parameter estimation method in Chapter III and further regularizing the negative log-likelihood in the outer step, we propose a two-step algorithm as follows.

- Inner step: Given $\{\theta(\cdot), X[0]\}$ or given η^*, we solve the ODE with time-varying coefficient by using the following criteria:

$$\hat{c}(\eta^*) = \arg\min_c \int \sum_{j=1}^{m} w_j \left[\frac{d\hat{X}_j}{dt} - F_j\{\hat{X}, \theta(\cdot), t\} \right]^2 dt \text{ subject to } \hat{X}[0] = X[0], \tag{4.3}$$

where w_j's are adjusted normalizing weights with the purpose of making the numerical magnitudes of different components comparable. In this step, we construct the estimated function $\hat{c}$ given η^*, which will be adopted into the outer step as follows.

- Outer step: We can estimate η^* by minimizing a penalized least square:

$$\hat{\eta}^* = \arg\min_{\eta^*} \sum_{j=1}^{m} \sum_{i=1}^{n_j} w_j \{Y_{ij} - \hat{X}_j(\eta^*, t_{ij})\}^2 + \sum_{\ell=1}^{d} \lambda_\ell \int_0^T |\theta'_\ell(\tau)| d\tau, \qquad (4.4)$$

  where Y_{ij} and t_{ij} are the i-th observed value and time respectively for the j-th ODE component, $\hat{X}_j$ is a function of η^* and estimated from the inner step and λ_ℓ is the tuning parameter for the ℓ-th coefficient curve. With the first derivative of the time-varying coefficients regularized by penalties in the outer step, we aim at "identifying" the regions where each coefficient curve remains a constant, respectively. When $\lambda_\ell \to \infty$ sufficiently fast, $\theta_\ell(t) \equiv \theta_\ell$, which is a constant across the whole study range.

The minimization problem in the inner step can be fastly solved with a non-linear least square step by providing the corresponding derivatives or gradients. As in Ramsay, et al. (2007), by applying the *implicit function theorem* we obtain the derivatives $\partial\hat{c}/\partial\eta$ and $\partial\hat{c}/\partial X[0]$, which can be used to construct the non-linear least square gradient in the outer step. Without having to handle the penalties for $\theta(\cdot)$ and the associating additional variation in the inner step of Ramsay, et al. (2007), the minimization problem in the outer step can be easily solved. Note that our approach also provide a balance between goodness of fit between the observations and the assumed parametric ODE system and the fidelity toward that ODE system by allowing part of $\theta(t)$ to be time-varying. Suppose we use K equally spaced points, $\tau_1, \cdots, \tau_K = T$, between 0 and T to approximate the integral in the penalty and rewrite the penalty as

$$P_\lambda(\theta) = \sum_{\ell=1}^{d} \lambda_\ell \bigtriangledown_\tau \sum_{k=1}^{K} |\theta'_\ell(\tau_k)|, \tag{4.5}$$

where $\bigtriangledown_\tau$ is the distance between two neighbor points τ_k and τ_{k+1}. We can absorb $\bigtriangledown_\tau$ into the tuning parameter λ_ℓ in numerical calculation and use the local quadratic approximation in Fan and Li (2001) to approximate the absolute functions in the above penalty. Lastly we can carry out the minimization problem in the outer step via non-linear least square. In practice, we use the same λ_ℓ to reduce computation cost. One can take large enough p to expand the ODE curves, but very large p may inflate estimation variation. For a given number of observations, the optimal value of p can be determined by the estimation rate from the following theoretical discussion. To penalize the derivative of ODE curve over the whole time range, K can be much larger than p.

4.2.2 Penalty Tuning Parameter Selection

The tuning parameter selection is commonly done using criteria such as Bayesian Information Criterion (BIC), Akaki Inofrmation Crierion (AIC) or generalized cross validation (GCV) methods, provided that the degree of freedom can be obtained. For an estimator $\hat{\mu}$, the conventional definition of the degree of freedom is

$$df = \sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i)/\sigma^2.$$

Efron et. al. (2004) proposed a bootstrap procedure and Shen and Ye (2002), a data perturbation method, to estimate the degree of freedom associating with $\hat{\mu}$. Even though these procedures can be calculated in a straightforward fashion, they are computationally demanding. To reduce the computation cost, we design an approximation to calculate the degree of freedom by mimicking how it is estimated in the ridge regression.

Denote $\boldsymbol{Y}_j = (Y_{1j}, \cdots, Y_{n_j j})^T$ and $\hat{\boldsymbol{X}}_j = \{\hat{X}_j(t_{1j}), \ldots, \hat{X}_j(t_{n_j j})\}^T = \boldsymbol{\Phi}_j c_j$ with $\boldsymbol{\Phi}_j$ $= \{\phi(t_{1j}), \cdots, \phi(t_{n_j j})\}^T$ being an $n_j \times q$ matrix. In (4.4), by conducting a first order Taylor's expansion at the estimated $\hat{\eta}^*$, we rewrtie the first term, the L_2-norm square, as

$$\sum_{j=1}^{m} w_j \left\| \boldsymbol{Y}_j - \boldsymbol{\Phi}_j \hat{c}_j|_{\hat{\eta}^*} - \boldsymbol{Z}_j(\eta^* - \hat{\eta}^*) \right\|^2$$

where $\boldsymbol{Z}_j = \boldsymbol{\Phi}_j \left.\frac{\partial \hat{c}_j}{\partial \eta^*}\right|_{\hat{\eta}^*}$. With η^* being the argument that we intend to maximize, the above approximation gives the regular least square component in the ridge regression. We then put the penalty term into the ridge structure. By applying a local quadratic approximation at $\hat{\eta}^*$, following by a second order Taylor's expansion at a part of $\hat{\eta}^*$, namely $\hat{\eta}_\ell$, we approximate $P_\lambda(\theta)$ in (4.5), or equivalently the second term in (4.4), by

$$\sum_{\ell=1}^{d} \lambda_\ell \sum_{k=1}^{K} \left[|\hat{\theta}'_\ell(\tau_k)| + \operatorname{sgn}\{\hat{\theta}'_\ell(\tau_k)\} \psi'(\tau_k)^T(\eta_\ell - \hat{\eta}_\ell) + (\eta_\ell - \hat{\eta}_\ell)^T \frac{\psi'(\tau_k)\psi'^T(\tau_k)}{2|\hat{\theta}'_\ell(\tau_k)|}(\eta_\ell - \hat{\eta}_\ell) \right].$$

Recall that $\eta^* = (X[0]^T, \eta^T)^T$ and write

$$\begin{aligned}
\boldsymbol{D} &= \operatorname{diag}\left\{ \lambda_1 \sum_{k=1}^{K} \frac{\psi'(\tau_k)\psi'^T(\tau_k)}{2|\hat{\theta}'_1(\tau_k)|}, \cdots, \lambda_d \sum_{k=1}^{K} \frac{\psi'(\tau_k)\psi'^T(\tau_k)}{2|\hat{\theta}'_d(\tau_k)|}, \underbrace{0, \cdots, 0}_{m} \right\} \\
\boldsymbol{Z} &= \left(Z_1^T, \cdots, Z_m^T \right)^T \\
\boldsymbol{Z}_w &= \left(w_1 Z_1^T, \cdots, w_m Z_m^T \right)^T
\end{aligned}$$

Mimicking the degree of freedom calculation within the ridge regression framework, we estimate the degree of freedom in ours by

$$df = \operatorname{Tr}\left[\boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z}_w + \boldsymbol{D})^{-1} \boldsymbol{Z}_w \right].$$

Let $\boldsymbol{V}$ be the vector consisting of the diagonal elements of $\boldsymbol{Z}(\boldsymbol{Z}^T \boldsymbol{Z}_w + \boldsymbol{D})^{-1} \boldsymbol{Z}_w$ and denote the degree of freedom within each ODE component by $df_j, j = 1, \ldots, m$, then $df_j = \sum_{i=n_1+\cdots+n_{j-1}+1}^{n_1+\cdots+n_j} \boldsymbol{V}_i$. Our numerical investigation shows that similar outcomes are obtained when estimating the degree of freedom either by bootstrap or by the above approximation. The gain in using the approximation is its computation speed. With the estimated degree of freedom, we select the tuning parameters using the following selection criteria:

$$\begin{aligned}
\text{BIC} &= \sum_{j=1}^{m} w_j \left(\frac{\|\boldsymbol{Y}_j - \hat{\boldsymbol{X}}_j\|^2}{\hat{\sigma}_j^2} + \log n_j \cdot df_j \right), \\
\text{AIC} &= \sum_{j=1}^{m} w_j \left(\frac{\|\boldsymbol{Y}_j - \hat{\boldsymbol{X}}_j\|^2}{\hat{\sigma}_j^2} + 2 \cdot df_j \right), \\
\text{GCV} &= \sum_{j=1}^{m} w_j \left(\frac{\|\boldsymbol{Y}_j - \hat{\boldsymbol{X}}_j\|^2}{\hat{\sigma}_j^2 (n_j - df_j)^2} \right),
\end{aligned}$$

where $\hat{\sigma}_j$ can be estimated through non-penalized method in the outer step.

4.3 Theoretical Results

In this section we study some non-asymptotic bounds of our proposed estimation method. We also derive convergence rates for our estimators under certain regularity conditions. The proofs are given in the Appendix. To clearly present the fundamental concepts, we focus on the scenario that there is only one time-varying coefficient; that is, $d = 1$, and the ODE system (4.1) has one X with $m = 1$. The total time range is also scaled to $[0, 1]$ in the proof; that is, $T = 1$.

Recall that we use an element in $\mathbb{L}_{\psi,p}$, space expanded by the basis functions $\psi(\tau)$, to approximate the ODE coefficient curve $\theta(\tau)$. In a vector format, we have, with each element of $\xi(\tau) \in \mathbb{L}_{\psi,p}$,

$$\theta(\tau) = \psi(\tau)^T \eta + e_p(\tau) = \xi(\tau) + e_p(\tau). \tag{4.6}$$

In this approximation, we let $\omega_p = \|\theta - \xi\|_\infty = \|e_p\|_\infty$, where $\|\cdot\|_\infty$ is the supremum norm.

With $0 < \tau_1 < \tau_2 < \cdots < \tau_K = 1$ forming a grid of K evenly spaced points between 0 and 1, we let $A = \{\bigtriangledown\psi(\tau_1), \bigtriangledown\psi(\tau_2), \cdots, \bigtriangledown\psi(\tau_K)\}^T$, where $\bigtriangledown\psi(\tau_k) = K\{\psi(\tau_k) - \psi(\tau_{k-1})\}$. Letting $\gamma_k = K\{\xi(\tau_k) - \xi(\tau_{k-1})\}$ and approximating the first derivative $\theta'(\tau_k)$ by γ_k, we have

$$\gamma = A\eta, \tag{4.7}$$

where $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_K)^T$. For exposition simplicity in our theoretical development and observing the connection between the coefficient functions and their derivatives, hereafter, we let $K = p$; consequently, A is a $p \times p$ square matrix. For certain choices of ψ, for example B-splines, A is also invertible. Denote $\gamma^* = (X[0], \gamma^T)^T$. Using ξ to approximate the non-parametric curve θ and replacing ξ equivalently by γ, we write the estimated ODE curve $\hat{X}$ in the inner step as $\hat{X}(\gamma^*, t)$ or $\hat{X}(\xi, X[0], t)$. That is, given the approximation of θ, ξ, and initial value $X[0]$, $\hat{X}(\xi, X[0], t)$ is the solution of (4.1).

To further simplify the setup, we first approximate the integral of the penalty with the $K = p$ evenly spaced points. We then rewrite the outer criterion (4.4) as

$$\hat{\gamma}^* = \arg\min_{\gamma^*} \frac{1}{n}\|\boldsymbol{Y} - \hat{X}(\gamma^*, t)\|^2 + \frac{\lambda}{p}\|\gamma\|_1, \tag{4.8}$$

where $\|\gamma\|_1 = \sum_{k=1}^p |\gamma_k|$, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ and $t = (t_1, t_2, \ldots, t_n)^T$.

We need the following assumptions for our theoretical outcomes.

(C1) Given $\xi(\tau)$, we assume that the force function F in the ODE system equation (4.1) satisfies

$$|F(x, \xi, \cdot) - F(y, \xi, \cdot)| \leq c_1|x - y|.$$

In the inner step, We use an element in a q-dimensional functional space to approximate the ODE curve. We add Assumption A2 to warrant the estimation precision.

(C2) For any ξ, there exist an element, $X_{[q]}(t)$, in a q-dimensional functional space, $\mathbb{L}_{\phi,q}$, such that the ODE curve $X(\xi, X[0], t)$ can be approximated by a function $X_{[q]}(t)$ and that

$$\begin{aligned} \|X_{[q]}(\cdot) - X(\xi, X[0], \cdot)\|_\infty &\leq r_q, \\ \left\| \frac{dX_{[q]}}{dt}(\cdot) - \frac{dX}{dt}(\xi, X[0], \cdot) \right\|_\infty &\leq r_q, \end{aligned}$$

where the rate of r_q is determined by the number of basis function, q, or the choice of knots when B-spline basis are applied to approximate the ODE curve. If cubic splines are used to approximate the ODE curve, the assumption A2 can be easily satisfied based on Theorem 1 in Hall (1968). We now introduce Lemma **??**.

Lemma IV.1. *Under Assumptions C1-C2, considering that $\hat{X}(\xi, X[0], \cdot)$ is estimated using the inner criterion (4.3), we have, given ξ and $X[0]$,*

$$\|\hat{X}(\xi, X[0], \cdot) - X(\xi, X[0], \cdot)\|_\infty \leq c_2 r_q, \tag{4.9}$$

where c_2 is a constant that is a function of c_1 in Assumption C1.

From this lemma, we note that, given parameter γ^*, the ODE curve can be estimated with a high precision. The precision rate is only determined by the estimation rate r_q in Assumption C2.

To ensure the estimation uniqueness in the inner step, i.e., the estimated ODE curve is uniquely determined by the vector of γ^*, and vice versa, we assume

(C3) There exist $\mathcal{C}_k^U(t)$ and $\mathcal{C}_k^L(t)$, $k = 1, \ldots, p+1$, such that for given γ^* and $\gamma^\dagger$, the estimated ODE curve satisfies

$$\sum_{k=1}^{p+1} \mathcal{C}_k^L(t)|\gamma_k^* - \gamma_k^\dagger| \leq |\hat{X}(\gamma^*, t) - \hat{X}(\gamma^\dagger, t)| \leq \sum_{k=1}^{p+1} \mathcal{C}_k^U(t)|\gamma_k^* - \gamma_k^\dagger| \tag{4.10}$$

with probability one. Furthermore, given initial value $X[0]$ and two coefficient functions, θ_1 and θ_2, we assume that the ODE curves satisfy $|X(\theta_1, X[0], t) - X(\theta_2, X[0], t)| \leq c_3\|\theta_1 - \theta_2\|_\infty$.

Let $\mathcal{M}^U$ and $\mathcal{M}^L$ be matrices with elements $(\mathcal{M}^U)_{ik} = \mathcal{C}_k^U(t_i)$ and $(\mathcal{M}^L)_{ik} = \mathcal{C}_k^L(t_i)$, $i = 1, \ldots, n$ and $k = 1, \ldots, p+1$. Denote $J_\mathcal{F} = \{2, \ldots, p+1\}$ and $J(\gamma) = \{k \in J_\mathcal{F} : \gamma_k \neq 0\}$, and $|J|$ denotes the cardinality of J. For a vector $\Delta \in \mathbb{R}^{p+1}$ and a subset $J \subset J_\mathcal{F}$, we let Δ_J have the same coordinates as Δ on J and zero coordinates on the rest of the entries. We also let Δ_J^* have the same coordinate as Δ on the kth entry, $k \in J \bigcup \{1\}$, and zero coordinates elsewhere. Similar to the restrictive eigenvalue assumption in Bickel, et al. (2009), the matrix $\mathcal{M}^L$ satisfies the following assumption.

(C4)

$$\kappa \equiv \min_{J \subseteq J_\mathcal{F}: |J| \leq s} \; \min_{\Delta \neq 0: \|\Delta_{J_\mathcal{F}}\|_1 \leq 4\|\Delta_J^*\|_1} \frac{\|\mathcal{M}^L \Delta\|}{\sqrt{n}\|\Delta_J^*\|} > 0, \tag{4.11}$$

where $1 \leq s \leq p$.

Define $K_1 = \bigvee_{k=2}^{p+1} \|\mathcal{M}_{\cdot k}^U\| \bigvee (4^{-1}\|\mathcal{M}_{\cdot 1}^U\|)$ and $K_2 = \bigvee_{k=1}^{p+1} (\|\mathcal{M}_{\cdot k}^U\|_1/n)$ where $a \vee b = \max(a, b)$, and $\mathcal{M}_{\cdot k}^U$ denotes the k-th column of $\mathcal{M}^U$. Denote $\alpha_p(\tau) = \|\psi(\tau)^T A^{-1}\|$ and $\omega_{p,q} = c_2 r_q + c_3 \omega_p$.

Theorem IV.2. *For the true ODE parameter curve $\theta_0(\cdot)$, we denote the approximated curve in $\mathbb{L}_{\psi,p}$ and the coefficient vector by $\xi_0(\cdot)$ and $\gamma_0^* = (X_0[0],\ \gamma_0^T)^T$. Define $S_p = |J(\gamma_0)|$. Let $Y_i = X(\theta_0, X[0], t_i) + \varepsilon_i = \hat{X}(\gamma_0^*, t_i) + \zeta_i$; $\zeta_i = \varepsilon_i + e_i$. Assume $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and let*

$$\frac{\lambda}{2p} = a\sigma_\varepsilon K_1 \sqrt{\frac{\log(p+1)}{n}} + 2K_2\omega_{p,q},$$

and $a > 2\sqrt{2}$. Then, under Assumptions C1-C4 and for all $n \geq 1$, with probability at *least* $1-(p+1)^{1-a^2/8}$ *we have*

$$\begin{aligned}|\hat{X}_0[0]-X_0[0]| &\leq \frac{4a\sigma_\varepsilon K_1}{\kappa^2}\sqrt{\frac{(S_p+1)\log(p+1)}{n}}+\frac{8K_2\sqrt{S_p+1}}{\kappa^2}\omega_{p,q},\\ |\hat{\theta}_0(\tau)-\theta_0(\tau)| &\leq \frac{16\alpha_p(\tau)a\sigma_\varepsilon K_1(S_p+1)}{\kappa^2}\sqrt{\frac{\log(p+1)}{n}}+\frac{32\alpha_p(\tau)K_2(S_p+1)}{\kappa^2}\omega_{p,q}+\omega_p.\end{aligned}$$

In addition, if $\omega_{p,q}=o[\{\log(p+1)/n\}^{1/2}]$ *so that approximately* $\zeta_i\sim N(0,\sigma_\varepsilon^2)$ *then we have the following error bounds:*

$$\begin{aligned}|\hat{X}_0[0]-X_0[0]| &\leq \frac{4a\sigma_\varepsilon K_1}{\kappa^2}\sqrt{\frac{(S_p+1)\log(p+1)}{n}},\\ |\hat{\theta}_0(\tau)-\theta_0(\tau)| &\leq \frac{16\alpha_p(\tau)a\sigma_\varepsilon K_1(S_p+1)}{\kappa^2}\sqrt{\frac{\log(p+1)}{n}}.\end{aligned}$$

The proof of the above theorem is given in Appendix C. From the conclusions in the above theorem, if, under certain conditions, $\alpha_p(\tau)$ and κ remain bounded as n and p diverge, the bounds are proportional to $\{\log(p+1)/n\}^{1/2}+\omega_{p,q}+\omega_p$, where the latter two terms declines to zero as p and q grow to ∞. Hence, the non-asymptotic results hold under that high-dimensional scenario when $p\gg n$. Further, as in the case when $\theta(t)\equiv\theta$ which is time invariant, we can safely use a very large q in approximation $X(t)$. This is because there is no bias-variance tradeoff in that estimation.

With additional Assumptions C5-C9 below, we can use the bounds presented in Theorem IV.2 to derive asymptotic convergence rates for $\hat{X}[0]$ and $\hat{\theta}_0(\tau)$. Their validity is mostly controlled by the choice of $\psi(\tau)$ and A.

(C5) There exists a constant $S<\infty$ such that $S_p\leq S$ for all p; further, K_1 and K_2 are both bounded.

(C6) There exists a constant $\nu>0$ such that $(p+1)^\nu r_q$ and $(p+1)^\nu\omega_p$ are bounded. Then $(p+1)^\nu\omega_{p,q}$ is also bounded.

(C7) For a given τ, there exists $b_\tau<\nu$ such that $(p+1)^{-b_\tau}\alpha_p(\tau)$ is bounded for all n

and p.

(C8) There exists $\mu < \nu$ such that $(p+1)^{-\mu} \sup_\tau \alpha_p(\tau)$ is bounded for all n and p.

(C9) κ is bounded away from zero for large enough n, where $n \to \infty$, $p \to \infty$ and $p/n \to 0$.

Corollary IV.3. *With Assumptions A5-A9, if we let p grow at the rate of $n^{1/(2\nu)}$, the estimators $\hat{X}[0]$ and $\hat{\theta}(\tau)$ in Theorem IV.2 satisfy the following converging rate with probability tending to one:*

$$\begin{aligned} |\hat{X}[0] - X_0[0]| &= O\left(\sqrt{\frac{\log n}{n}}\right), \\ |\hat{\theta}(\tau) - \theta^0(\tau)| &= O\left(\frac{\sqrt{\log n}}{n^{1/2-b_\tau/(2\nu)}}\right), \\ \sup_\tau |\hat{\theta}(\tau) - \theta^0(\tau)| &= O\left(\frac{\sqrt{\log n}}{n^{1/2-\mu/(2\nu)}}\right). \end{aligned}$$

Furthermore, if $\zeta_i \sim N(0, \sigma_\varepsilon^2)$ and we let p grow at the rate of $n^{1/(2\nu+2b_\tau)}$, the converging rate of $\hat{\theta}(\tau)$ becomes

$$\begin{aligned} |\hat{X}[0] - X_0[0]| &= O\left(\sqrt{\frac{\log n}{n}}\right) \\ |\hat{\theta}(\tau) - \theta^0(\tau)| &= O\left(\frac{\sqrt{\log n}}{n^{\nu/(2\nu+2b_\tau)}}\right), \end{aligned}$$

with probability tending to one while the converging rate of $\hat{X}[0]$ stays same. If we let p grow at the rate of $n^{1/(2\nu+2\mu)}$, we have the following supremum converging rate for $\hat{\theta}$:

$$\sup_\tau |\hat{\theta}(\tau) - \theta^0(\tau)| = O\left(\frac{\sqrt{\log n}}{n^{\nu/(2\nu+2\mu)}}\right).$$

We provide the proof of the above corollary in Appendix C for the completeness. From these results, one can easily find that the converge rate for $\hat{X}[0]$ is slightly better than the non-parametric estimator $\hat{\theta}(\tau)$ in all cases.

4.4 Simulation Study

In this section, we apply the proposed method to two simulated ODE dynamic systems; both have wide scientific applications.

Exmaple 1. We first study the FitzHugh-Nagumo ODE model. This model was invented by FitzHugh (1961) and Nagumo et al. (1962) to simply the Hodgkin-Huxley model (1952), which was used to study the behavior of spike potential in the giant axon of squid neurons. The ODE equations are

$$\frac{dV}{dt} = c\left(V - \frac{V^3}{3} + R\right), \ \frac{dR}{dt} = -\frac{1}{c}(V - a + bR), \tag{4.12}$$

where V describes the voltage across an axon membrane, R is the recovery variable summarizing outward currents, and $a,\ b,\ c$ are ODE coefficient parameters in the dynamic system. In this simulation study, we let the parameters $a,\ b$ and c in (4.12) vary over time according to the following settings:

$$\begin{aligned} a &= 0.2\mathbb{I}(0 \le t \le 3) + \frac{1}{45}[18 - (t-6)^2]\mathbb{I}(3 < t \le 9) + 0.2\mathbb{I}(9 < t \le 20), \\ b &= 0.2\mathbb{I}(0 \le t \le 11) + \frac{1}{45}[18 - (t-14)^2]\mathbb{I}(11 < t \le 17) + 0.2\mathbb{I}(17 < t \le 20), \\ c &= 3.0\mathbb{I}(0 \le t \le 7) + \frac{1}{6}[9 + (t-10)^2]\mathbb{I}(10 < t \le 13) + 3.0\mathbb{I}(13 < t \le 20), \end{aligned}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The observations of V and R from (4.12) are simulated with initial values $V(0) = -1.0$ and $R(0) = 1.0$. We generate 201 pairs of V and R with the observation errors from $N(0, 0.5^2)$ on equally-distanced grids in [0.20]. The cubic B-splines with knots placed at each observation time are used to approximate the ODE curves. We use the cubic B-splines with 21 equally spaced knots for estimating the ODE coefficient curves. In the penalty term, the 201 τ_k 's with an equal distance between two consecutive points are used to approximate the integral.

For each simulated data, we consider four estimation methods, Method I-IV. Method I is the proposed method that regularizes the outer-step criteria. Method II is the varying-coefficient estimation method without the regularization penalties. In Method III, the parametric model with constant parameters are considered. In addition to these three method, we also compare the results with Method IV in which the true constant regions of parameters are known. In true constant regions, the parameters are estimated as constant values, and cubic B-splines are only applied in non-constant regions. Method IV can only be carried out in the simulation study. For Method I, we also compare the results using, respectively, BIC, AIC and GCV criteria for tuning parameter selection.

Denote the true ODE coefficient curves by $\theta_\ell^0(\tau), \ell = 1, \ldots, d$. We let $\boldsymbol{F}_\ell = \{\tau : \theta_\ell^{0\prime}(\tau) = 0\}$ and $\boldsymbol{F}_\ell^c = \{\tau : \theta_\ell^{0\prime}(\tau) \neq 0\}$. In the example, θ^0 are a, b and c. Let $|\boldsymbol{F}_\ell|$ and $|\boldsymbol{F}_\ell^c|$ be the lengths of the regions of $\boldsymbol{F}_\ell$ and $\boldsymbol{F}_\ell^c$ respectively. The following mean integrated square errors (MISE) are used to compare different methods

$$\text{MISE}_{\boldsymbol{F}_\ell} = \text{E}\int_{\boldsymbol{F}_\ell} [\hat{\theta}_\ell(\tau) - \theta_\ell^0(\tau)]^2 d\tau, \ \text{MISE}_{\boldsymbol{F}_\ell^c} = \text{E}\int_{\boldsymbol{F}_\ell^c} [\hat{\theta}_\ell(\tau) - \theta_\ell^0(\tau)]^2 d\tau.$$

The MISE is reported using the sample mean over 100 repetitions. Denote the standard error of the estimator $\hat{\theta}_\ell$ at τ as $\text{SE}_{\hat{\theta}_\ell}(\tau)$ and define the average estimation standard errors (AVSE) as

$$\text{AVSE}_{\boldsymbol{F}_\ell} = |\boldsymbol{F}_\ell|^{-1} \int_{\boldsymbol{F}_\ell} \text{SE}_{\hat{\theta}_\ell}(\tau) d\tau, \ \text{AVSE}_{\boldsymbol{F}_\ell^c} = |\boldsymbol{F}_\ell^c|^{-1} \int_{\boldsymbol{F}_\ell^c} \text{SE}_{\hat{\theta}_\ell}(\tau) d\tau.$$

For the estimation of the ODE curves $X_j(t)$, which are $V(t)$ and $R(t)$ in this simulation, we also compare the MISE and AVSE for estimated X_j and they are defined as follows:

$$\text{MISE}_{X_j} = \text{E}\int_0^T [\hat{X}_j(t) - X_j(t)]^2 dt,\ \text{AVSE}_{X_j} = T^{-1}\int_0^T \text{SE}_{\hat{X}_j}(t)dt,$$

where $\hat{X}_j$ is the the estimated curve of the j-th ODE component X_j and $\text{SE}_{\hat{X}_j}(t)$ denotes the standard error of $\hat{X}_j$ at time t.

The MISE results over 100 repetitions for both the parameter curves and the ODE curves are reported in the top half of Table 1. The AVSE results are also compared in the bottom half of Table 1. The average estimated parameter curves are plotted with the true cures in Figure 1. From the MISE results, our method performs the best among the four methods considered. For Method I, the BIC results are slightly better than those of AIC and GCV. The parameter curves are constant over most of time and BIC reduces the model complexity in terms of first derivatives much more than the other two. The performance of Method II is worst among all methods, specially for the MISE results of a and b. For c, Method II performs slightly better, particularly in the region of $\boldsymbol{F}_c^c$. This observation implies that the parameter c is more sensitive to the structure of the FitzHugh-Nagumo dynamic system and is easier to be estimated than a and b. It further suggests that c can not be estimated as a constant over all time, as also indicated by the MISE results of Method III. In the comparisons of AVSE results among all four methods, because the penalties reduce the model complexity, our proposed Method I has comparable standard errors with Method III, in which only five parameters, i.e., constant a, b, c and two initial values, need to be estimated. Even though it seems that Method IV outperforms Method III in Figure 1, the MISE results within the constant regions of $\boldsymbol{F}_a$ and $\boldsymbol{F}_b$ suggest otherwise. The reason is that Method III gains by the smaller estimation standard errors. From Fig. 1, we can find that in the non-constant time regions $\boldsymbol{F}_a^c$, $\boldsymbol{F}_b^c$ and $\boldsymbol{F}_c^c$. The proposed Method I over-shrinks the bump and valley, specially at the points where the true first derivatives are zero. The reason is due to the penalties of the first derivative for the parameter curves. From the estimation results of ODE curves V and R, Method I with BIC achieves the smallest MISE results and the smallest AVSE results, and Method III performs the worst.

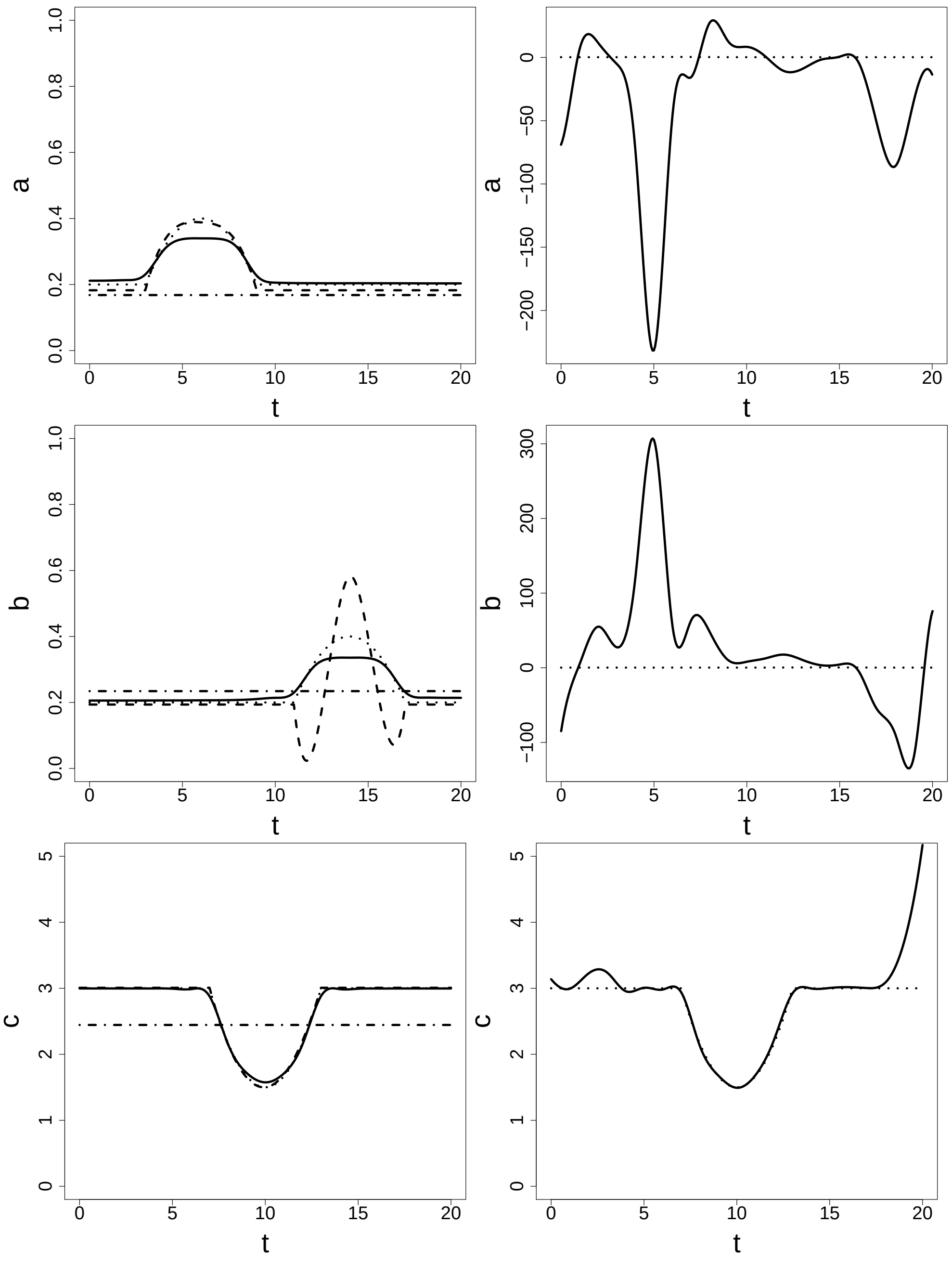

Figure 4.1: Estimated parameter cures plotted with true curves in the FitzHugh-Nagumo model. Left half of the figure: Method I (solid); Method III (dash-dot); Method VI (dashed); True (dotted). Right of the figure: Method II (solid); True (dotted).

		MISE results						
	$\boldsymbol{F}_a$	$\boldsymbol{F}_a^c$	$\boldsymbol{F}_b$	$\boldsymbol{F}_b^c$	$\boldsymbol{F}_c$	$\boldsymbol{F}_c^c$	V	R
BIC	0.0022	0.0096	0.0036	0.0106	0.0045	0.0166	0.0717	0.0112
AIC	0.0126	0.0160	0.0433	0.0174	0.0100	0.0292	0.1028	0.0274
GCV	0.0093	0.0151	0.0311	0.0159	0.0091	0.0275	0.1000	0.0245
Method II	273.8949	27.1258	742.8140	18.5210	126.0361	0.9516	0.1262	0.0569
Method III	0.0238	0.1894	0.0920	0.1125	4.3427	2.3924	2.7103	0.7336
Method IV	0.2749	0.5590	0.5189	1.4269	0.6427	0.9301	0.1158	0.0498

		AVSE results						
	$\boldsymbol{F}_a$	$\boldsymbol{F}_a^c$	$\boldsymbol{F}_b$	$\boldsymbol{F}_b^c$	$\boldsymbol{F}_c$	$\boldsymbol{F}_c^c$	V	R
BIC	0.0143	0.0193	0.0195	0.0243	0.0161	0.0191	0.0341	0.0197
AIC	0.0298	0.0311	0.0453	0.0347	0.0159	0.0391	0.0459	0.0325
GCV	0.0265	0.0301	0.0390	0.0313	0.0142	0.0361	0.0446	0.0305
Method II	3.7451	4.4388	9.4972	3.0448	1.2656	0.5976	0.0570	0.0474
Method III	0.0262	0.0262	0.0738	0.0738	0.0370	0.0370	0.0614	0.0418
Method IV	0.2494	0.3318	0.3093	0.6889	0.3204	0.5639	0.0568	0.0485

Table 4.1: The average MISE and AVSE results of the four different estimation methods over 100 repetitions in the FitzHugh-Nagumo model.

Exmaple 2. In the second example, the well known Lotka-Volterra dynamic model (Lotka 1910 and Volterra 1926) are studied. This model, also known as predator-prey model, has wide applications in modeling the dynamics of ecological systems with predator-prey interactions, competition and disease dispersion. The model has two components H and L, described by the following ODEs:

$$\frac{dH}{dt} = aH - bHL, \ \frac{dL}{dt} = -c + dHL,$$

where a, b, c and d are the ODE parameters, and they are given by the following equations in the simulation:

$$\begin{aligned}
a &= 0.5\mathbb{I}(0 \le t \le 6) + \frac{1}{30}[2.4 - (t-9)^2]\mathbb{I}(6 < t \le 12) + 0.5\mathbb{I}(12 < t \le 20), \\
b &= 0.25\mathbb{I}(0 \le t < 3) + \frac{1}{90}[13.5 + (t-6)^2]\mathbb{I}((3 < t \le 9) + 0.25\mathbb{I}((9 < t \le 20), \\
c &= 1.0\mathbb{I}(0 \le t < 11) + \frac{1}{45}[36 + (t-14)^2]\mathbb{I}(11 < t \le 17) + 1.0\mathbb{I}(17 < t \le 20), \\
d &= 0.3\mathbb{I}(0 \le t \le 6) + \frac{1}{45}[22.5 - (t-9)^2]\mathbb{I}(6 < t \le 12) + 0.3\mathbb{I}(12 < t \le 20)(4.13)
\end{aligned}$$

We generate 101 pairs of observations still on equally-distanced grid with the initial values $H(0) = 3.5$ and $L(0) = 0.5$. The observation errors are from $N(0,1)$. Knots are placed at each observed time point and the corresponding cubic B-splines are used to approximate the two ODE curves. The setting of cubic B-splines and the grid for approximation of the penalty integral are the same as those in Example 1.

Performances of Method I - IV are also investigated for this simulation example. The MISE and AVSE results over 100 time simulations are shown in Table 4.2, and the average estimation parameter curves are plotted in Figure 4.2. From Figure 4.2, the estimation results of Method II are not so much as erratic as the estimation results of a and b in the previous example. It means that the parameters in this example are easier to be estimated, compared to those in the previous FitzHugh-Nagumo ODE model.

The average estimated parameter curves are very close to the true curves for both Method I and Method IV. From MISE and AVSE results in Table 4.2, our propose Method I still performs best among the four methods. Different from the results of the previous example, the MISE and AVSE results of Method III are unanimously worse than other Methods because Method III barely fits the data simulated using our parameter setting (4.13). From the very large MISE and AVSE results of H and L obtained from Method III in Table 4.2, we can also find this reason.

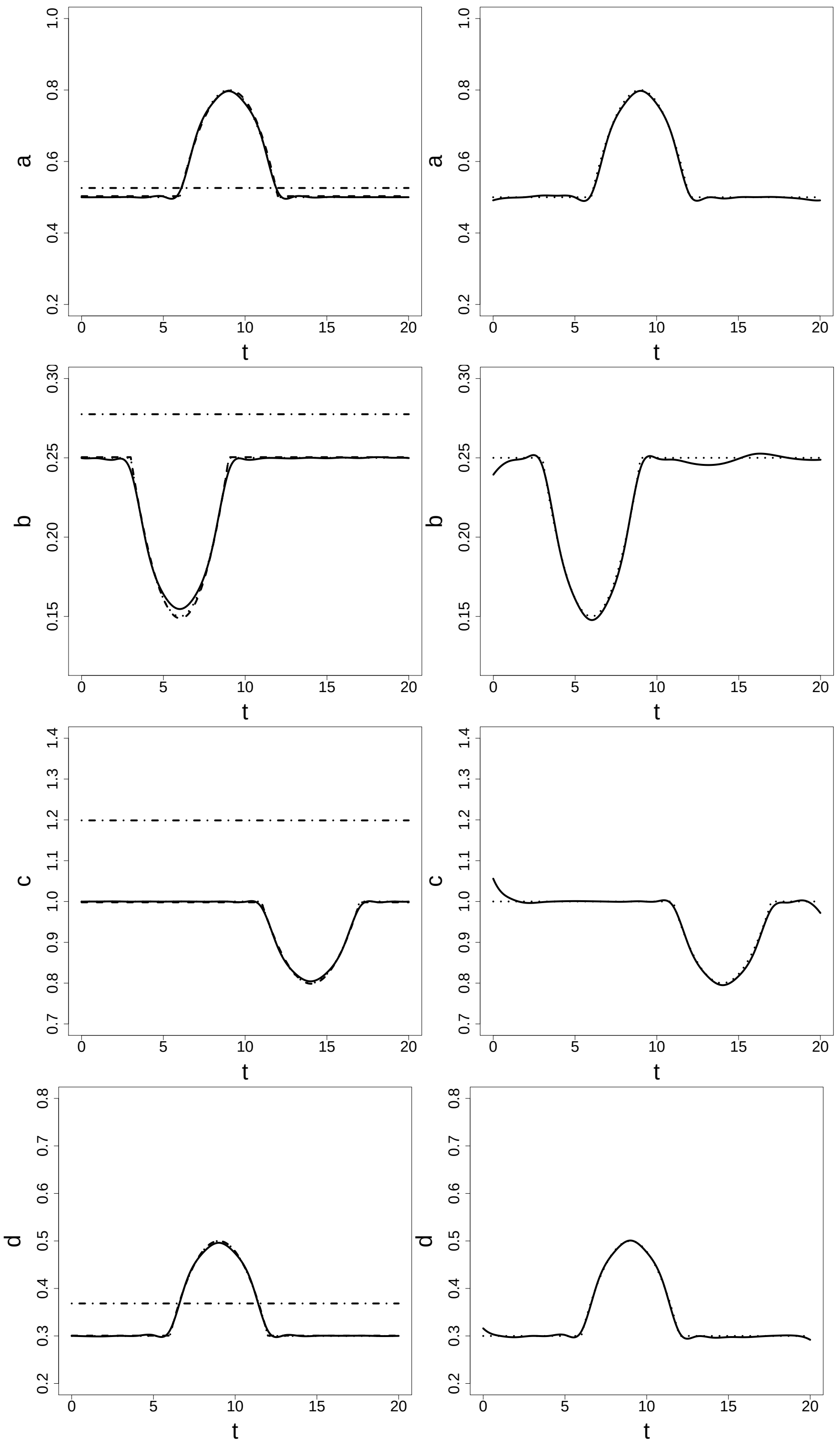

Figure 4.2: Estimated parameter cures plotted with true curves in the Lotka-Volterra model. Lines are as in Fig. 1.

		MISE results								
	$\boldsymbol{F}_a$	$\boldsymbol{F}_a^c$	$\boldsymbol{F}_b$	$\boldsymbol{F}_b^c$	$\boldsymbol{F}_c$	$\boldsymbol{F}_c^c$	$\boldsymbol{F}_d$	$\boldsymbol{F}_d^c$	H	L
BIC	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.4451	0.4602
AIC	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.6902	0.7596
GCV	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.6578	0.7818
Method II	1.9567	0.1343	0.9532	0.0030	0.2355	0.0254	0.0185	0.0004	0.7205	0.8989
Method III	0.2043	0.2880	0.0210	0.0341	0.6938	0.6967	0.0713	0.0548	65.5308	82.5574
Method IV	0.0001	0.0027	0.0002	0.0002	0.0003	0.0004	0.0002	0.0001	0.6483	0.8114

		AVSE results								
	$\boldsymbol{F}_a$	$\boldsymbol{F}_a^c$	$\boldsymbol{F}_b$	$\boldsymbol{F}_b^c$	$\boldsymbol{F}_c$	$\boldsymbol{F}_c^c$	$\boldsymbol{F}_d$	$\boldsymbol{F}_d^c$	H	L
BIC	0.0011	0.0033	0.0016	0.0035	0.0010	0.0031	0.0017	0.0012	0.0916	0.1128
AIC	0.0011	0.0033	0.0016	0.0034	0.0010	0.0033	0.0018	0.0013	0.0921	0.1135
GCV	0.0011	0.0032	0.0016	0.0033	0.0010	0.0033	0.0018	0.0013	0.0923	0.1137
Method II	0.0280	0.0161	0.0250	0.0041	0.0087	0.0069	0.0107	0.0023	0.1718	0.1904
Method III	0.2354	0.2354	0.0805	0.0805	0.3853	0.3853	0.1054	0.1054	1.5541	1.4706
Method IV	0.0052	0.0212	0.0069	0.0061	0.0076	0.0099	0.0081	0.0087	0.1636	0.1797

Table 4.2: The average MISE and AVSE results of the four different estimation methods over 100 repetitions in the Lotka-Volterra model.

4.5 Analysis of Ecology and Epidemiological Data Sets

1. Hare and lynx data: We first apply our proposed method to a lynx-hare data set with a predator-prey dynamic model. The numbers of trapped lynx and snowshoe hares of North Canada were collected from 1900 to 1920 (Odum 1953). The observed data should reflect the relative populations of lynx and hare in the study region. The dynamic ODE model follows as Lotka-Volterra dynamic model (4.13) with H and L representing the evolution function of the number of snowshoe hares and lynx respectively. The cubic B-splines with 201 equally-spaced knots are used to estimate the ODE curves of $H(t)$ and $L(t)$. The parameter curves of a, b, c and d are estimated with the cubic B-splines but only with 8 equally-spaced knots due to the limited number of observation.

We use the BIC criterion to regularize the ODE coefficient curves, and plot them together with the estimated curves by non-regularized method (Method II) and that of a constant fitting (Method III) in Figure 4.3. With the data set only containing 21 pair observations, we almost fail to obtain any structure information for ODE parameters by Method II. With the penalties, Method I gives us improved structural information over that of Method II: the estimated parameter a stays almost constant over the whole observation time period; the estimated parameter b is basically a constant from 1900 to 1905, then starts to increase and later stabilizes into a slightly larger constant around 1912; the estimated parameter c stays constant over the whole time period; the estimated parameter d varies quite a bit and has not reached a constant-stage across the whole time, but it stills has much smaller variation in comparison to the estimated d curve of Method II.

From Figure 4.4, both Method I and II fit better at the first peaks of H and L dynamics. Method III performs worse than the other two at the the middle valleys of H and L. Looking at the tails of $H(t)$ and $L(t)$, we conclude that Method I performs the best among all three methods. A bootstrap analysis was conducted for a variation comparison. The result, which is not reported here, shows that standard errors of our regularized method are comparable to those of Method III, and are much smaller than that standard errors of Method II.

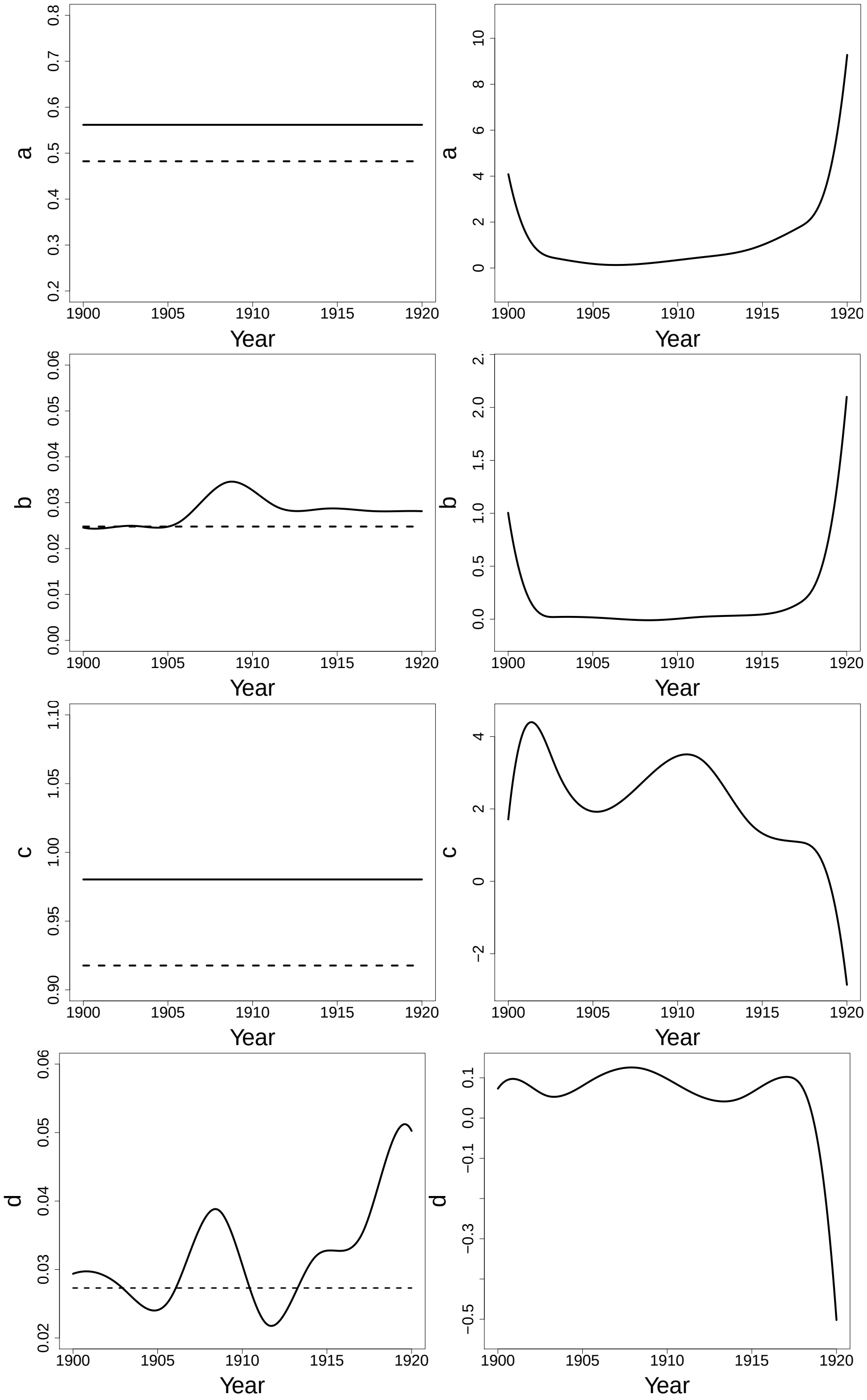

Figure 4.3: Estimated parameter cures for the lynx-hare data set through the Lotka-Volterra model. Left half of the figure: Method I (solid); Method III (dashed); Right half of the figure: Method II (solid).

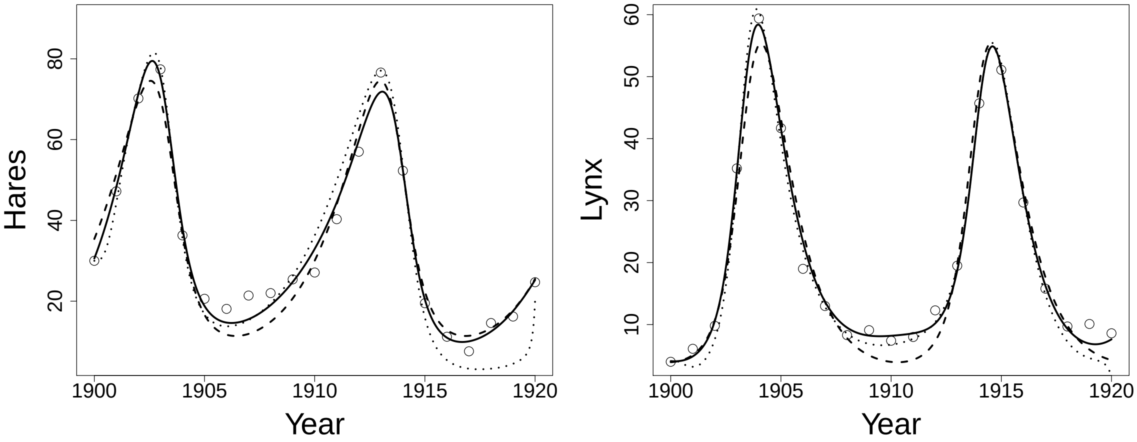

Figure 4.4: Estimated $H(\cdot)$ and $L(\cdot)$ for the lynx-hare data set. Method I (solid); Method II (dotted); Method III (dashed).

2. Canadian Measles Incidence Dynamic data: In the second data example, we apply our method to analyze a Canadian Measles incidence dynamic data set. The data consists of weekly measles incidence reports for the province of Ontario, Canada, from 1939 through 1965. A scatter plot of the data is given in Hooker, et al. (2011). Following their analysis, the data can be modeled through the so called SEI dynamic equations:

$$\begin{aligned}\dot{S} &= \rho(t) - \beta(t)\left(\frac{I}{p(t)} + v\right)S, \\ \dot{E} &= \beta(t)\left(\frac{I}{p(t)} + v\right)S - \sigma E, \\ \dot{I} &= p(t)\sigma E - \gamma I,\end{aligned}$$

where S is the susceptible class, E is the exposed (infected with the disease but not infectious) class and I is the infectious class. S increases with a recruitment rate $\rho(t)$ and moves into E with a rate of $\beta(t)(I/p(t) + v)$. E transforms into I with the rate $\sigma p(t)$ as I recovers with the rate γ. In this data set only I, the measles infectious class, is observed. The other two state variable S and E are unobserved. The parameters are $\rho(t)$, $\beta(t)$, $p(t)(= p_0 + p_1(t - 1952))$, v, σ and γ. $\rho(t)$ is interpolated from the monthly birth rate data at a five-year lag. σ is known to be around 8 days, i.e., $\sigma = 365/8$. γ is roughly estimated by the five day mean infectious period and equals to 365/5. Only the parameters $\beta(t)$, p_0, p_1 and v need to be estimated from the data.

The structure of $\beta(t)$ within each year has been studied in Bauch and Earn (2003), which consists of a high-level component at the summer season and a low-level one during the rest of the year. Adopting this yearly structure, we further use the proposed method to find the long-term pattern of $\beta(t)$. Following Hooker, et al. (2011), we let

$$\beta(t) = \alpha(t) + \theta(t) \tag{4.14}$$

where $\theta(t)$ is a cyclic function that describes the same within-year pattern across all years, which is subject to the constrain $\int_0^1 \theta(t)dt = 0$, and $\alpha(t)$ is the general coefficient function that describes the long-term time trend. We use the cyclic cubic B-splines with knots on each month to expand $\theta(t)$ while using the regular cubic B-splines with knots on each year for approximating $\alpha(t)$. We only regularize $\alpha(t)$ to find the long-term yearly pattern.

We compare our results with two other methods: the varying-coefficient approach without regularization penalty and a set of short-term constant-fit conducted every two years. In latter, we assume $\alpha(t) = c$ for two neighboring years and only fit the data within that two years and repeat this process for about 25 times from 1939 to 1963. In Figure 4.5, the regularized and non-regularized $\alpha(t)$ are plotted with the two-year constant fitting, all in log scale. We find that the regularized $\alpha(t)$ is larger at the early years and deceases gradually to a constant after 1958. This means the rate at which the susceptible class moving into the exposed class decreases in the long-term pattern and gradually becomes stabilized. This pattern of $\alpha(t)$ could be due to an introduction of measles vaccine around 1954. After the measles vaccine took effect, $\alpha(t)$ could be modeled as a constant and $\beta(t)$ only contains the seasonal pattern. To give a simple illustration, we also plot the fitting results for I in Figure 4.5 from 1952 to 1954. In term of the prediction error, the two-year constant fitting approach perform the best but it is not straight forward to learn the pattern from this modeling strategy. Our regularized method performs slightly better than the non-regularized approach.

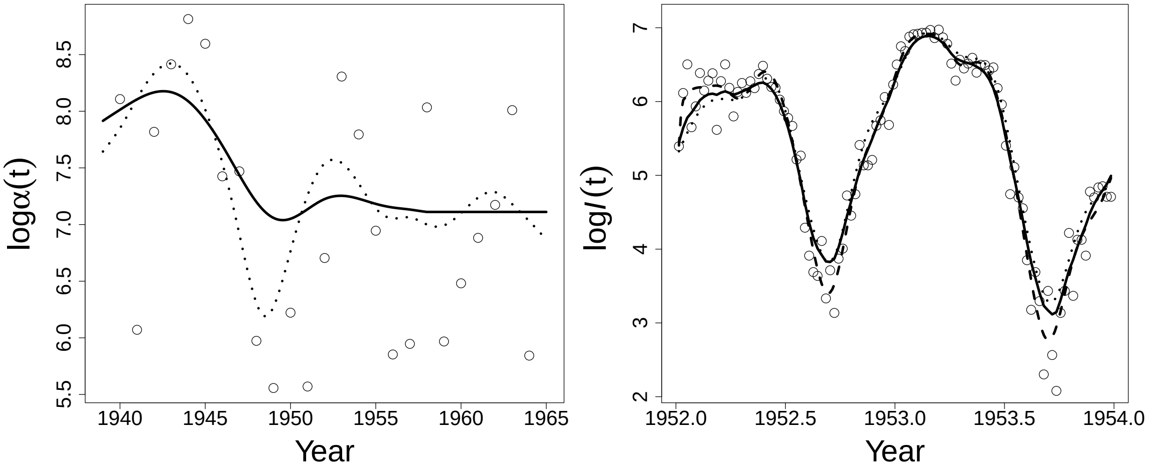

Figure 4.5: Left: estimated $\alpha(t)$ for regularized method (solid) and non-regularized method (dotted); circles are for the two-year constant fitting. Right: estimated $I(t)$ for 1952-1954; regularized method (solid), non-regularized method (dotted) and two-year constant fitting (dashed).

4.6 Discussion

We proposed a regularized parameter estimation method for the ODE dynamic system. If the ODE coefficient curve has parts of derivatives being 0, i.e., constant in some regions in time, the regularized approach in general performs much better than the ordinary non-regularized method. Under this situation, one can also utilize the parametric ODE structure with new independent errors to predict the range of new observations in the near future. To this end, prediction can be conducted, as in the stochastic scenario, by heavily relying on the assumed the model being correct and possessing a parametric structure. The penalties not only help in recovering the parametric structure, but also play a role of smoothing and on reducing the estimation variation. From our theoretical study, the nonparametric curves obtain estimation bounds as functions of $\{\log p\}^{1/2}$ under certain regularity conditions. This indicates that a large p may not cause much harm in prediction accuracy for the purposed method.

APPENDICES

APPENDIX A

Convex Regularization Method for High-Dimensional Grouped Variable Selection

We first prove the following lemma before the proof of Theorem II.3

Lemma A.1. *Consider the model* (2.2)*. Assume that the random variables* $\varepsilon_1, \cdots, \varepsilon_n$ *are i.i.d. normal with mean zero and variance* σ^2*, and all diagonal elements of the matrix* $X^T X/n$ *are equal to 1. Suppose* $M_G(\beta^*) = s$ *and* $M(\beta^*) = r$*. Let*

$$\lambda_2 = \rho\lambda_1 = 2\rho A\sigma\sqrt{\frac{\log p}{n}},$$

and $A > \sqrt{2}$*. Then with probability at least* $1 - p^{1-A^2/2}$*, for any solution* $\hat{\beta}$ *of minimization problem* (2.17) *and all* $\beta \in \mathbb{R}^p$ *we have*

$$\begin{aligned} & \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 + 2\lambda_2\sum_{k=1}^{K}\|\hat{\beta}^k - \beta^k\| + \lambda_1|\hat{\beta}-\beta| \\ \leq \; & \frac{1}{n}\|X(\beta-\beta^*)\|^2 + 4\lambda_2\sum_{k\in J_G(\beta)}\|\hat{\beta}^k-\beta^k\| + 4\lambda_1\sum_{kj\in J(\beta)}|\hat{\beta}_{kj}-\beta_{kj}|, \quad \text{(A.1)} \\ M(\hat{\beta}) \leq \; & \frac{4}{\lambda_1^2 n^2}\|X^T X(\hat{\beta}-\beta^*)\|^2. \quad \text{(A.2)} \end{aligned}$$

Proof $\forall \beta \in \mathbb{R}^p$, we have

$$\frac{1}{n}\|X\hat{\beta} - y\|^2 + 2\lambda_2 \sum_{k=1}^{K} \|\hat{\beta}^k\| + 2\lambda_1|\hat{\beta}| \leq \frac{1}{n}\|X\beta - y\|^2 + 2\lambda_2 \sum_{k=1}^{K} \|\beta^k\| + 2\lambda_1|\beta|. \quad \text{(A.3)}$$

By using $y = X\beta^* + \varepsilon$, the above inequality is equivalent to

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 \leq \frac{1}{n}\|X(\beta-\beta^*)\|^2 + \frac{2}{n}\varepsilon^T X(\hat{\beta}-\beta) + 2\lambda_2 \sum_{k=1}^{K}(\|\beta^k\| - \|\hat{\beta}^k\|) + 2\lambda_1(|\beta| - |\hat{\beta}|). \quad \text{(A.4)}$$

Note that

$$\varepsilon^T X(\hat{\beta} - \beta) \leq |X^T\varepsilon|_\infty |\hat{\beta} - \beta|_1, \quad \text{(A.5)}$$

where $|X^T\varepsilon|_\infty = \max_{1\leq k\leq K, 1\leq j\leq L} |\sum_{i=1}^{n} X_{i,kj}\varepsilon_i|$. We consider the random event

$$\mathcal{A} = \left\{ \frac{2}{n}|X^T\varepsilon|_\infty \leq \lambda \right\}. \quad \text{(A.6)}$$

Since all diagonal elements of the matrix $X^T X/n$ are equal to 1, the following random variables

$$V_{kj} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_{i,kj}\varepsilon_i \ 1 \leq k \leq K, \ 1 \leq j \leq L, \quad \text{(A.7)}$$

are i.i.d. standard normal. Using this fact we can obtain the following probability equalities for any kj

$$Pr(|\sum_{i=1}^{n} X_{i,kj}\varepsilon_i)| \geq \frac{\lambda_1 n}{2}) = Pr(|Z| \geq \frac{\lambda_1\sqrt{n}}{2\sigma}), \quad \text{(A.8)}$$

where Z is the standard normal random variable. By applying the tail bound on the normal distribution and plugging in $\lambda_1 = 2A\sigma\sqrt{\frac{\log p}{n}}$, we have the probability upper bound for the event $\mathcal{A}^c$:

$$\begin{aligned} Pr(\mathcal{A}^c) &\leq p \cdot Pr(|Z| \geq \frac{\lambda_1\sqrt{n}}{2\sigma}), \\ &\leq p^{1-A^2/2}. \end{aligned} \tag{A.9}$$

Following (A.4), on the event $\mathcal{A}$, we have

$$\begin{aligned} & \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 + 2\lambda_2\sum_{k=1}^{K}\|\hat{\beta}^k-\beta^k\| + \lambda|\hat{\beta}-\beta| \\ \leq\ & \frac{1}{n}\|X(\beta-\beta^*)\|^2 + 2\lambda_2\sum_{k=1}^{K}(\|\hat{\beta}^k-\beta^k\| + \|\beta^k\| - \|\hat{\beta}^k\|) + 2\lambda_1(|\hat{\beta}-\beta| + |\beta| - |\hat{\beta}|), \\ \leq\ & \frac{1}{n}\|X(\beta-\beta^*)\|^2 + 4\lambda_2\sum_{k\in J_G(\beta)}\|\hat{\beta}^k-\beta^k\| + 4\lambda_1\sum_{kj\in J(\beta)}|\hat{\beta}_{kj}-\beta_{kj}|. \end{aligned} \tag{A.10}$$

This is the first inequality in this lemma. To prove the second one, we first state the KKT conditions on the convex analysis of model (2.17):

$$\begin{cases} \frac{1}{n}(X^T(Y-X\hat{\beta}))_{kj} = \lambda_2\frac{\hat{\beta}_{kj}}{\|\hat{\beta}^k\|} + \lambda_1\mathrm{sgn}(\hat{\beta}_{kj}) & \forall\hat{\beta}_{kj}\neq 0, \\ \frac{1}{n}|(X^T(Y-X\hat{\beta}))_{kj}| \leq \lambda_1+\lambda_2 & \forall\hat{\beta}_{kj}=0. \end{cases} \tag{A.11}$$

Then we prove the inequality for the sparsity of $\hat{\beta}$, $M(\hat{\beta})$. From KKT condition and the definition of $\mathcal{A}$, it is not difficult to find that

$$\lambda_1 \leq \frac{1}{n}|(X^T(Y-X\hat{\beta}))_{kj}| \leq \frac{1}{n}|(X^TX(\hat{\beta}-\beta^*))_{kj}| + \frac{1}{2}\lambda_1 \text{ if } \hat{\beta}_{kj}\neq 0. \tag{A.12}$$

Therefore, we have

$$M(\hat{\beta}) \leq \frac{4}{\lambda_1^2n^2}\sum_{kj\in J(\hat{\beta})}|(X^TX(\hat{\beta}-\beta^*))_{kj}|^2 \leq \frac{4}{\lambda_1^2n^2}\|X^TX(\hat{\beta}-\beta^*)\|^2.$$

This completes the proof of the lemma.

Proof of Theorem II.3 The proof will follow the basic idea of Bickel et al. (2009) for the lasso method. Since the penalty in our method includes both L_1 and L_2 norms, the KKT conditions and many of the technical details will be different from that of lasso.

Let $J = J(\beta^*)$ and $J_G = J_G(\beta^*)$. By (A.1) in Lemma 2, with $\beta = \beta^*$, we have, on the event $\mathcal{A}$, that

$$\begin{aligned} \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 &\leq 4\lambda_2 \sum_{k\in J_G} \|\hat{\beta}^k - \beta^{k*}\| + 4\lambda_1 \sum_{kj\in J} |\hat{\beta}_{kj} - \beta^*_{kj}|, \\ &\leq 4\lambda_2\sqrt{s}\|(\hat{\beta}-\beta^*)_{J_G}\| + 4\lambda_1\sqrt{r}\|(\hat{\beta}-\beta^*)_J\|. \qquad \text{(A.13)} \end{aligned}$$

Moreover by the same inequality, we have

$$2\lambda_2 \sum_{k=1}^{K} \|\hat{\beta}^k - \beta^{k*}\| + \lambda_1|\hat{\beta}-\beta^*| \leq 4\lambda_2 \sum_{k\in J_G} \|\hat{\beta}^k - \beta^{k*}\| + 4\lambda_1 \sum_{kj\in J} |\hat{\beta}_{kj} - \beta^*_{kj}|,$$

which is equivalent to

$$2\lambda_2 \sum_{k\in J_G^c} \|\hat{\beta}^k - \beta^{k*}\| + \lambda_1 \sum_{kj\in J^c} |\hat{\beta}_{kj} - \beta^*_{kj}| \leq 2\lambda_2 \sum_{k\in J_G} \|\hat{\beta}^k - \beta^{k*}\| + 3\lambda_1 \sum_{kj\in J} |\hat{\beta}_{kj} - \beta^*_{kj}|.$$

This is the condition in Assumption 3.1 at $\Delta = \hat{\beta} - \beta^*$ since $\lambda_2/\lambda_1 = \rho$. Thus, by the assumption, we have

$$\begin{aligned} \|(\hat{\beta}-\beta^*)_{J_G}\| &\leq \frac{\|X(\hat{\beta}-\beta^*)\|}{\kappa_G\sqrt{n}}, \\ \|(\hat{\beta}-\beta^*)_J\| &\leq \frac{\|X(\hat{\beta}-\beta^*)\|}{\kappa\sqrt{n}}. \qquad \text{(A.14)} \end{aligned}$$

Plugging the above inequalities into (A.13), we have

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 \leq 16\lambda_1^2 \left(\frac{\rho\sqrt{s}}{\kappa_G} + \frac{\sqrt{r}}{\kappa}\right)^2. \qquad \text{(A.15)}$$

Also by noting that

$$2\rho \sum_{k=1}^{K} \|\hat{\beta}^k - \beta^{*k}\| + |\hat{\beta} - \beta^*| \leq 4\rho\sqrt{s}\|(\hat{\beta} - \beta^*)_{J_G}\| + 4\sqrt{r}\|(\hat{\beta} - \beta^*)_J\|,$$

and $\|\hat{\beta} - \beta^*\|_{2,1} \leq |\hat{\beta} - \beta^*|$, we have

$$\|\hat{\beta} - \beta^*\|_{2,1} \leq \frac{16\lambda_1}{2\rho + 1} \left(\frac{\rho\sqrt{s}}{\kappa_G} + \frac{\sqrt{r}}{\kappa} \right)^2. \tag{A.16}$$

By plugging the value of λ_1 into (A.15) and (A.16), we prove (2.19) and (2.20).

From (A.2) in Lemma 6.1, we obtain:

$$M(\hat{\beta}) \leq \frac{4}{\lambda_1^2 n^2} \|X^T X(\hat{\beta} - \beta^*)\|^2 \leq \frac{4\phi_{\max}}{\lambda_1^2 n} \|X(\hat{\beta} - \beta^*)\|^2.$$

By plugging the bound of $\|X(\hat{\beta} - \beta^*)\|^2$ into the above inequality, we have (2.21).

Finally, we need to prove (2.22). For the sake of notational simplicity, we let $\Delta = \hat{\beta} - \beta^*$ and let J'_G be the set of indices in J_G^c corresponding to s maximal in absolute value norms $\|\Delta^k\|$, and write $J_{G,2s} = J_G \cup J'_G$. Also similarly, let J' be the set of indices in J^c corresponding to r maximal in absolute value $|\Delta_{kj}|$, and write $J_{2r} = J \cup J'$. Note that $|J_{G,2s}| \leq 2s$ and $|J_{2r}| \leq 2r$. Denote $\|\Delta_{J_G^c}^{(l)}\|$ and $\|\Delta_{J^c}^{(l)}\|$ as the l-th largest value in the sets $\{\|\Delta^k\| : k \in J_G^c\}$ and $\{\|\Delta^{kj}\| : kj \in J^c\}$, respectively. We have

$$\begin{aligned} \|\Delta_{J_G^c}^{(l)}\| &\leq \sum_{k \in J_G^c} \|\Delta^k\|/l = \|\Delta_{J_G^c}\|_{2,1}/l, \\ \|\Delta_{J^c}^{(l)}\| &\leq \sum_{kj \in J^c} |\Delta_{kj}|/l = |\Delta_{J^c}|/l. \end{aligned} \tag{A.17}$$

Then we obtain

$$\begin{aligned}
\|\Delta_{J_{G,2s}^c}\|^2 &= \sum_{k \in J_{G,2s}^c} \|\Delta^k\|^2 \le \sum_{l=s+1}^{\infty} \frac{\|\Delta_{J_G^c}\|_{2,1}^2}{l^2} \le \frac{\|\Delta_{J_G^c}\|_{2,1}^2}{s}, \\
\|\Delta_{J_{2r}^c}\|^2 &= \sum_{kj \in J_{2r}^c} |\Delta_{kj}|^2 \le \sum_{l=r+1}^{\infty} \frac{|\Delta_{J^c}|^2}{l^2} \le \frac{|\Delta_{J^c}|^2}{r}, \qquad \text{(A.18)}
\end{aligned}$$

by the assumption RE(r, s, ρ), which implies that

$$\begin{aligned}
2\rho\sqrt{s}\|\Delta_{J_{G,2s}^c}\| + \sqrt{r}\|\Delta_{J_{2r}^c}\| &\le 2\rho\|\Delta_{J_G}\|_{2,1} + 3|\Delta_J|, \\
&\le 2\rho\sqrt{s}\|\Delta_{J_G}\| + 3\sqrt{r}\|\Delta_J\|, \\
&\le 2\rho\sqrt{s}\|\Delta_{J_{G,2s}}\| + 3\sqrt{r}\|\Delta_{J_{2r}}\|. \qquad \text{(A.19)}
\end{aligned}$$

Since

$$\begin{aligned}
\|\Delta\| &\le \|\Delta_{J_{G,2s}^c}\| + \|\Delta_{J_{G,2s}}\| \text{ and} \\
\|\Delta\| &\le \|\Delta_{J_{2r}^c}\| + \|\Delta_{J_{2r}}\|, \qquad \text{(A.20)}
\end{aligned}$$

we have

$$\|\Delta\| \le \frac{4\rho\sqrt{s}}{2\rho\sqrt{s} + \sqrt{r}}\|\Delta_{J_{G,2s}}\| + \frac{4\sqrt{r}}{2\rho\sqrt{s} + \sqrt{r}}\|\Delta_{J_{2r}}\|. \qquad \text{(A.21)}$$

In addition, the condition in the assumption RE(r, s, ρ) implies the following condition for the assumption RE$(2r, 2s, \rho)$:

$$|\Delta_{J_{2r}^c}| + 2\rho\|\Delta_{J_{G,2s}^c}\|_{2,1} \le 3|\Delta_{J_{2r}}| + 2\rho\|\Delta_{J_{G,2s}}\|_{2,1}.$$

Similar to (A.15), we have the following upper bound

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|^2 \leq 16\lambda_1^2\left(\frac{\rho\sqrt{2s}}{\kappa_G(2r,2s,\rho)}+\frac{\sqrt{2r}}{\kappa(2r,2s,\rho)}\right)^2. \tag{A.22}$$

Then by the additional assumption RE$(2r, 2s, \rho)$, we can obtain the upper bounds for $\|\Delta_{J_{G,2s}}\|$ and $\|\Delta_{J_{2r}}\|$, and then from (A.21), we have (2.22). This completes the proof. □

APPENDIX B

Parameter Estimation for Ordinary Differential Equations

Proof OF Proposition III.1

Recall that $\alpha = \lambda^{-1}$. Replacing X by $\hat{X}_a = \Phi c$, we let

$$J(c) = J(c, \lambda) = \alpha \|y - \Phi c\|^2 + \int \left(\frac{d\Phi}{dt} c - a \right)^2 dt.$$

From $\partial J(c)/\partial c = 0$, we obtain $\hat{c} = (\alpha \Phi^T \Phi + \int \Phi'^T \Phi' dt)^{-1} (\alpha \Phi^T y + a \int \Phi'^T dt)$. Plugging $\hat{c}$ into $H(a) = \|Y - \hat{X}_a\|^2$ and by letting $\partial H(a)/\partial a = 0$, we find $\hat{a} = (b^T b)^{-1} b^T (I - \alpha D) y$, where

$$b = b(\alpha) = \Phi(\alpha \Phi^T \Phi + \int \Phi'^T \Phi' dt)^{-1} \int \Phi'^T dt, \text{ and } D = D(\alpha) = \Phi(\alpha \Phi^T \Phi + \int \Phi'^T \Phi' dt)^{-1} \Phi^T.$$

Suppose $\mathrm{var}(y) = \sigma^2$, we have

$$\mathrm{var}(\hat{a}) = \frac{b^T (I - \alpha D)^2 b}{(b^T b)^2} \sigma^2.$$

For a matrix M,

$$\frac{\partial M^{-1}}{\partial \alpha} = -M^{-1} \frac{\partial M}{\partial \alpha} M^{-1}.$$

Consequently,

$$\frac{\partial}{\partial \alpha}\mathrm{var}(\hat{a}) = 4\frac{b^T D_\alpha b}{(b^T b)^3} b^T (I - \alpha D)\left(I - \frac{b^T b}{b^T D b} D\right)(I - \alpha D) b\sigma^2.$$

With $\alpha D = \Phi(\Phi^T \Phi + \alpha^{-1}\int \Phi'^T \Phi' dt)^{-1}\Phi^T$, the eigenvalues of αD satisfy $0 \leq \lambda_p \leq \lambda_{p-1} \leq \cdots \leq \lambda_1 \leq 1$. Denote by γ_i, δ_i, $i = 1, \cdots, p$, the eigenvalues of $I - \alpha D$ and $I - \frac{b^T b}{b^T D b} D$ respectively. Then we have

$$0 \leq \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_p \leq 1; \delta_1 \leq \delta_2 \leq \cdots \leq \delta_p.$$

According to the same non-descending order, the two sets of eigenvalues share the same set of eigenvectors $v_1, v_2, \cdots, v_p$. Since

$$b^T \left(I - \frac{b^T b}{b^T D b} D\right) b = 0,$$

we have

$$b^T \sum_{i=1}^{p} \delta_i v_i v_i^T b = \sum_{i=1}^{p} \delta_i (b_i^T v_i)^2 = 0.$$

Suppose $\delta_1 \leq \cdots \leq \delta_m \leq 0 \leq \delta_{m+1} \leq \cdots \leq \delta_p$, we have

$$\sum_{i=1}^{m} (-\delta_i)(b^T v_i)^2 = \sum_{j=m+1}^{p} \delta_i (b^T v_i)^2.$$

Therefore,

$$\begin{aligned}
& b^T(I-\alpha D)\left(I-\frac{b^Tb}{b^TDb}D\right)(I-\alpha D)b \\
= \;& b^T\sum_{i=1}^{p}\gamma_i v_i v_i^T \sum_{j=1}^{p}\delta_j v_j v_j^T \sum_{k=1}^{p}\gamma_k v_k v_k^T b \\
= \;& b^T\sum_{i=1}^{p}\gamma_i^2\delta_i v_i v_i^T b \\
= \;& \sum_{i=1}^{p}\gamma_i^2\delta_i (b^Tv_i)^2 \\
\geq \;& -\gamma_{m+1}^2\sum_{i=1}^{m}(-\delta_i)(b^Tv_i)^2+\gamma_{m+1}^2\sum_{i=m+1}^{p}\delta_i(b^Tv_i)^2=0.
\end{aligned}$$

This completes the proof.

Derivation of variance approximation for the one-component compartmental model:

Let

$$J(c)=\alpha\|y-\Phi c\|^2+\int\left(\frac{d\Phi}{dt}c-a\Phi c\right)^2dt.$$

From $\partial J(c)/\partial c=0$, we obtain $\hat{c}=\alpha M^{-1}(a,\alpha)\Phi^Ty$, where

$$M(a,\alpha)=\alpha\Phi^T\Phi+\int(\Phi'-a\Phi)^T(\Phi'-a\Phi)dt.$$

Plugging $\hat{c}$ into $H(a)=\|y-\hat{X}_a\|^2$ and by letting $\partial H(a)/\partial a=0$, we find

$$y^T(I-\alpha\Phi M^{-1}\Phi^T)\Phi M^{-1}(2a\int\Phi^T\Phi dt-\int\Phi^T\Phi' dt-\int\Phi'^T\Phi dt)M^{-1}\Phi^Ty=0.$$

Let

$$G(a,\alpha)=y^T\Lambda(a,\alpha)y,$$

where $\Lambda(a,\alpha)=(I-\alpha\Phi M^{-1}\Phi^T)\Phi M^{-1}(2a\int\Phi^T\Phi dt-\int\Phi^T\Phi' dt-\int\Phi'^T\Phi dt)M^{-1}\Phi^T$.

Perform Taylor expansion on $G(a,\alpha)$ around $a=a_0$, the true value of parameter a, then

$$0 = G(\hat{a}, \alpha) = G(a_0, \alpha) + \frac{\partial G}{\partial a}\bigg|_{a_0} (\hat{a} - a_0) + O_p(\hat{a} - a_0)^2.$$

We can approximate $\hat{a}$ as follows:

$$\hat{a} \approx a_0 - \left(\frac{\partial G(a_0, \alpha)}{\partial a} \right)^{-1} G(a_0, \alpha) = a_0 - \frac{y^T \Lambda(a_0, \alpha) y}{y^T \frac{\partial \Lambda(a_0,\alpha)}{\partial a} y}.$$

Then the variance of estimator $\hat{a}$ can be approximated as

$$\text{var}(\hat{a}) = \text{var} \left(\frac{y^T \Lambda(a_0, \alpha) y}{y^T \frac{\partial \Lambda(a_0,\alpha)}{\partial a} y} \right).$$

Using the realization of Y being normally distributed with the standard deviation $\sigma = 2$ and sample size $n = 201$, we calculated the values that were used to produce Figure 1.

Proof of Proposition III.2

Let

$$J(c|\alpha) = \alpha \|y - \Phi c\|^2 + \int \left\{ \frac{d\Phi}{dt} c - F(\Phi c, \theta^*) \right\}^2 dt,$$

and let $\hat{c}(\theta^*, \alpha)$ be the argument c that minimizes $J(c|\alpha)$ for given θ^* and α. From $\partial J(c|\alpha)/\partial c^T = 0$, we have $\hat{c}(\theta^*, \alpha)$ solves

$$\left\{ \alpha \Phi^T \Phi c + 1/2 \frac{\partial J(c|\alpha = 0)}{\partial c^T} \right\}\bigg|_{c = \hat{c}(\theta^*, \alpha)} = \alpha \Phi^T y. \tag{B.1}$$

Let $\widehat{c}_{\theta^*}(\theta^*, \alpha)$ and $\widehat{c}_\alpha(\theta^*, \alpha)$ denote the partial derivatives of $\widehat{c}(\theta^*, \alpha)$ with respect to α and θ^* respectively; and equivalently denote other partial derivatives; also let $J_{cc}(c|0)$ denote $\partial^2 J(c|0)/\partial c^T \partial c$. Direct derivations following taking derivatives of both sides of (B.1) show that (i) $\widehat{c}_{\theta^*}(\theta^*, \alpha)$ is free of the response Y, and (ii)

$$\left\{ \alpha \Phi^T \Phi + 1/2 J_{cc}(c|0) \right\} \widehat{c}_\alpha(\theta^*, \alpha) = \Phi^T \{ y - \Phi \widehat{c}_\alpha(\theta^*, \alpha) \}. \tag{B.2}$$

Hereafter, we need to assume that $\widehat{\theta}^* - \theta^*$ and α goes to zero. The derivations in the previous materials apply to all positive α and not just the ones near zero. Plugging $\hat{c}(\theta^*, \alpha)$ into $H\{c(\theta^*)\} = \|y - \hat{X}_{\theta^*}\|^2$, letting $\partial H\{c(\theta^*)\}/\partial\theta^{*T} = 0$, and deriving a Taylor series expansion on α at zero and $\widehat{\theta}^*$ at θ^*, we obtain

$$0 = G(\widehat{\theta}^*, \alpha) = \left\{G(\theta^*, 0) + G_{\theta^*}(\theta^*, \alpha)(\widehat{\theta}^* - \theta^*) + G_\alpha(\widehat{\theta}^*, 0)\alpha\right\}\{1 + o_p(1)\},$$

where

$$G(\theta^*, \alpha) = \frac{\partial \widehat{c}^T(\theta^*, \alpha)}{\partial \theta^*}\frac{\partial H(c)}{\partial c^T}|_{c=\widehat{c}(\theta^*,\alpha)} = -2\widehat{c}^T_{\theta^*}(\theta^*, \alpha)\Phi^T\{y - \Phi\widehat{c}(\theta^*, \alpha)\}. \qquad \text{(B.3)}$$

Thus, by noting G_{θ^*} is continuous at α near zero, we have

$$(\widehat{\theta}^* - \theta^*) \approx G^{-1}_{\theta^*}(\theta^*, 0)\left\{G(\theta^*, 0) + \alpha G_\alpha(\widehat{\theta}^*, 0)\right\}. \qquad \text{(B.4)}$$

Denoting $\left\{\alpha\Phi^T\Phi + 1/2J_{cc}(c|0)\right\}$ in (B.2) by $M_{c\alpha}$, assuming its inverse exists and deriving the structure of $G_\alpha(\widehat{\theta}^*, 0)$ using (B.3), we express $\{G(\theta^*, 0) + \alpha G_\alpha(\widehat{\theta}^*, 0)\}$ in (B.4) as

$$\begin{aligned} & -2d(\Phi, \widehat{c}, \theta^*, \alpha)\{y - \Phi\widehat{c}(\theta^*, 0)\}, \text{ where} & \text{(B.5)} \\ d(\Phi, \widehat{c}, \theta^*, \alpha) \quad &= \quad \widehat{c}^T_{\theta^*}\Phi^T + \alpha\left\{\widehat{c}^T_{\theta*\alpha} + \widehat{c}^T_{\theta*}\Phi^T\Phi M^{-1}_{c\alpha}\right\}\Phi^T. & \text{(B.6)} \end{aligned}$$

Further, by a direct application of semiparametric efficiency theory, it is well known that the most efficient expression for $d(\Phi, \widehat{c}, \theta^*, \alpha)$ in (B.5) is $\partial\Phi\widehat{c}(\theta^*, 0)/\partial\theta^*$, which is obtained when α is exact zero in (B.6). This efficiency statement does not require the errors to be normaly distributed. The smallest variance of $\widehat{\theta}^*$ in (B.4) is obtained when $\alpha = 0$ or equivalently by dropping the log-likelihood term in the inner step.

Assumptions for asymptotic properties:

Assumption B1. The density $q(t)$ of Q on $[0,T]$ is bounded by two positive numbers u, l, i.e., $u \leq q(t) \leq l$.

Assumption B2. For F, each component F_i, $i = 1, \cdots, m$, $\in C^3(\mathbb{R}^m \times [0,T] \times \Theta)$. For each θ and initial values $X(0)$, there exists a unique solution $X(\theta^*, t)$ of the (1.1) on $[0,T]$ and for any $\theta^* \neq \theta^{*'}$, we have $X(\theta^*, t) \neq X(\theta^{*'}, t)$.

Assumption B3. In $\Theta \times \Gamma$, define

$$M(\theta^*) = E_{\theta_0^*}\left[\sum_{i=1}^{m} w_i \ell_i(Y_i(t), X_i(\theta^*, t))\right],\ \theta^* \in (\Theta \times \Gamma)$$

where $\ell_i(Y_i(t), X_i(\theta^*, t))$ is the negative log-likelihood of the each i-th component observation (Y_i, X_i) at time t. We assume that $M(\theta^*)$ is continuous and has a unique maximum at θ_0^*

Assumption B4. For all i, $\ell_i(y, x)$ is a function in $C(\mathbb{R} \times \mathbb{R})$. If the random variable Y_i is not bounded, we assume that for any compact set $\Lambda \subset \mathbb{R}$,

$$\lim \inf_{|y| \to \infty} \left[\frac{-1 + \inf_{x \in \Lambda} \ell_i(y, x)}{\sup_{x \in \Lambda} \ell_i(y, x)}\right] > 0$$

Assumption B5. For all i, $\ell_i(y, x)$ is a function in $C^1(\mathbb{R} \times \mathbb{R})$ with

$$E_{\theta_0^*}\left[\left|\frac{\partial \ell_i}{\partial x}(Y_i(t), X_i(\theta_0^*, t))\right|\right] < \infty$$

If Y_i is not bounded, we assume that for any compact set $\Lambda \subset \mathbb{R}$,

$$\lim \inf_{|y| \to \infty} \left[\frac{-1 + \inf_{x \in \Lambda} \ell_i(y, x)}{\sup_{x \in \Lambda} \ell_i(y, x)}\right] > 0 \text{ and } \lim \inf_{|y| \to \infty} \left[\frac{1 + \inf_{x \in \Lambda} |\partial \ell_i / \partial x(y, x)|}{\sup_{x \in \Lambda} |\partial \ell_i / \partial x(y, x)|}\right] > 0.$$

Assumption B6. For all i, $\ell_i(y, x)$ is a function in $C^2(\mathbb{R} \times \mathbb{R})$ with

$$E_{\theta_0^*}\left[\left|\frac{\partial \ell_i}{\partial x}(Y_i, X_i(\theta_0^*, T))\right|^2\right] < \infty \text{ and } E_{\theta_0^*}\left[\left|\frac{\partial^2 \ell_i}{\partial x^2}(Y_i, X_i(\theta_0^*, T))\right|\right] < \infty$$

If Y_i is not bounded, we assume that for any compact set $\Lambda \subset \mathbb{R}$,

$$\lim_{|y|\to\infty} \inf \left[\frac{-1+\inf_{x\in\Lambda} \ell_i(y,x)}{\sup_{x\in\Lambda} \ell_i(y,x)}\right] > 0,\ \lim_{|y|\to\infty} \inf \left[\frac{1+\inf_{x\in\Lambda} |\partial\ell_i/\partial x(y,x)|}{\sup_{x\in\Lambda} |\partial\ell_i/\partial x(y,x)|}\right] > 0.$$

and

$$\lim_{|y|\to\infty} \inf \left[\frac{1+\inf_{x\in\Lambda} |\partial^2\ell_i/\partial x^2(y,x)|}{\sup_{x\in\Lambda} |\partial^2\ell_i/\partial x^2(y,x)|}\right] > 0.$$

<u>Proof of Lemma III.4:</u> Note that on the compact set $\Theta_0 \times \Gamma_0$, continuous functions F_i are bounded and so are the functions v_i in B_n. Since F_i has continuous partial derivatives, we can find a positive constant K on the bounded compact set, such that

$$|F_i(X,t,\theta) - F_i(X',t,\theta)| \ \leq \ K\sum_{j=1}^m w_j|X_j - X_j'|.$$

Define

$$J = J(X,\theta) = \sum_{i=1}^m w_i \int \left(\frac{dX_i}{dt} - F_i(X,t,\theta)\right)^2 dt$$

Given any $\theta^* \in \Theta_0 \times \Gamma_0$, and suppose $v = (v_1, \cdots, v_m)^T \in B_n^m$. From the definition of $\hat{X}_{ni}(\theta^*, \cdot)$, we have

$$\begin{aligned}
J(\hat{X}_n,\theta) &\leq J(v,\theta) \\
&\leq \sum_{i=1}^m w_i \int (\frac{dv_i}{dt} - F_i(v,t,\theta))^2 dt \\
&= \sum_{i=1}^m w_i \int (\frac{dv_i}{dt} - \frac{dX_i}{dt} + F_i(X,t,\theta) - F_i(v,t,\theta))^2 dt \\
&\leq \sum_{i=1}^m 2w_i \int (\frac{dv_i}{dt} - \frac{dX_i}{dt})^2 dt + 2w_i \int (F_i(X,t,\theta) - F_i(v,t,\theta))^2 dt \\
&\leq 2\sum_{i=1}^m w_i \int (\frac{dv_i}{dt} - \frac{dX_i}{dt})^2 dt + 8K^2 w_i \int (v_i - X_i)^2 dt \text{ by (C.1)} \\
&\leq 2T(\sum_{i=1}^m w_i \|\frac{dv_i}{dt} - \frac{dX_i}{dt}\|_\infty^2 + 4K^2 w_i(\|v_i - X_i\|_\infty^2) \\
&\leq 4mT(8K^2+2)r_n^2, \text{ by the definition of } r_n.
\end{aligned}$$

By the definition of J, we have

$$\sup_{\theta^* \in \Theta_0 \times \Gamma_0} w_i \int (\frac{d\hat{X}_{ni}}{dt} - F_i(\hat{X}_n, t, \theta))^2 dt \leq 4mT(8K^2 + 2) r_n^2.$$

Therefore,

$$\sup_{\theta^* \in \Theta_0 \times \Gamma_0} w_i \| \hat{X}_{ni} - X_i(0) - \int\limits_0^t F_i(\hat{X}_n, s, \theta) ds \|_\infty \leq T\sqrt{4m(8K^2 + 2)} r_n$$

Define

$$\begin{aligned} A_{ni}(\theta^*, t) &\equiv w_i \left(\hat{X}_{ni} - X_i(0) - \int\limits_0^t F_i(\hat{X}_n, s, \theta) ds \right) \\ &= w_i \left((\hat{X}_{ni} - X_i) - \int\limits_0^t (F_i(\hat{X}_n, s, \theta) - F_i(X, s, \theta)) ds \right) \end{aligned}$$

Then

$$\begin{aligned} w_i |\hat{X}_{ni} - X_i| &\leq w_i \int\limits_0^t |F_i(\hat{X}_n, s, \theta) - F_i(X, s, \theta)| ds + |A_{ni}(\theta^*, t)| \\ &\leq w_i K \sum_{j=1}^m \int\limits_0^t w_j |\hat{X}_{nj} - X_j| ds + \sup_{\theta^* \in \Theta_0 \times \Gamma_0} \|A_n(\theta^*, \cdot)\|_\infty \end{aligned}$$

By summing the above inequality over i, we have

$$\sum_{i=i}^m w_i |\hat{X}_{ni} - X_i| \leq mK \int\limits_0^t \sum_{j=1}^m w_j |\hat{X}_{nj} - X_j| ds + \sum_{i=1}^m \sup_{\theta^* \in \Theta_0 \times \Gamma_0} \|A_{ni}(\theta^*, \cdot)\|_\infty$$

By using Gronwall's inequality, we have

$$\sum_{i=i}^m w_i |\hat{X}_{ni} - X_i| \leq \sum_{i=1}^m \sup_{\theta^* \in \Theta_0 \times \Gamma_0} \|A_{ni}(\theta^*, \cdot)\|_\infty e^{mKT}$$

By the infinity norm bounds of A_{ni}, we have

$$\sup_{\theta^* \in \Theta_0 \times \Gamma_0} \sum_{i=1}^{m} w_i \|\hat{X}_{ni} - X_i\|_\infty \leq mT\sqrt{4m(8K^2+2)}r_n e^{mKT}$$

This completes the proof.

<u>Proof of Theorems III.5 and III.6:</u> The outer step in the proposed approach is the same as that of the RHCC method in Qi and Zhao (2010). When $r_n = o_p(1/n)$, the upper-bound given in Lemma III.4 ensures that for all θ^*, the differences between $X(\theta^*, t)$ and its estimates bear a negligible rate. This is equivalent to what Qi and Zhao have established for the inner-step of the RHCC estimates. Consequently, one can follow the exact steps in the proofs of Theorem 3.2 and 3.3 in Qi and Zhao (3.3) to prove the consistency and normality for the proposed estimator $\hat{\theta}^*_n$. □

APPENDIX C

Regularized Semiparametric Estimation for Ordinary Differential Equations

Proof of Lemma IV.1: The proof of this lemma mainly follows that of Theorem 3.1 in Qi and Zhao (2010) in which $\theta(t)$ is a constant. From Assumption C3, we have

$$|F(x,\xi,t) - F(x',\xi,t)| \ \leq \ c_1|x - x'|. \tag{C.1}$$

Define

$$J = J(X,\xi) = \int \left\{ \frac{dX}{dt} - F(X,\xi,t) \right\}^2 dt$$

Suppose that $v \in \mathbb{L}_{\phi,q}$ is an estimator of X. From the definition of $\hat{X}(\cdot)$, we have

$$\begin{aligned}
J(\hat{X},\xi) &\leq J(v,\xi) \\
&\leq \int (\frac{dv}{dt} - \frac{dX}{dt} + F(X,\xi,t) - F(v,\xi,t))^2 dt \\
&\leq 2\int (\frac{dv}{dt} - \frac{dX}{dt})^2 dt + 2\int (F(X,\xi,t) - F(v,\xi,t))^2 dt \\
&\leq 2\int (\frac{dv}{dt} - \frac{dX}{dt})^2 dt + 2c_1^2 \int (v - X)^2 dt \text{ by (C.1)} \\
&\leq 2T(\|\frac{dv}{dt} - \frac{dX}{dt}\|_\infty^2 + c_1^2(\|v - X\|_\infty^2) \\
&\leq 2T(c_1^2 + 1)r_q^2, \text{ by the definition of } r_q.
\end{aligned}$$

By the definition of J, we have

$$\int (\frac{d\hat{X}}{dt} - F(\hat{X}, \xi, t))^2 dt \leq 2T(c_1^2 + 1)r_q^2.$$

Hence using the same derivations behind Equation (1.3) in Qi and Zhao (2010), we have

$$\|\hat{X} - X[0] - \int_0^t F(\hat{X}, \xi, s)ds\|_\infty \quad \leq \quad T\sqrt{2(c_1^2 + 1)}r_q$$

Define

$$\begin{aligned} A(\xi, X[0], t) &\equiv \hat{X} - X[0] - \int_0^t F(\hat{X}, \xi, s)ds \\ &= (\hat{X} - X) - \int_0^t \{F(\hat{X}, \xi, s) - F(X, \xi, s)\}ds \end{aligned}$$

Then

$$\begin{aligned} |\hat{X} - X| &\leq \int_0^t |F(\hat{X}, \xi, s) - F(X, \xi, s)|ds + |A(\xi, X[0], t)| \\ &\leq c_1 \int_0^t |\hat{X} - X|ds + \|A(\xi, X[0], \cdot)\|_\infty \end{aligned}$$

By using Gronwall's inequality, we have

$$|\hat{X} - X| \leq \|A(\xi, X[0], \cdot)\|_\infty e^{c_1 T}$$

From the infinity norm bound of A, we have

$$\|\hat{X} - X\|_\infty \leq \left\{T\sqrt{2(c_1^2 + 1)}e^{c_1 T}\right\} r_q$$

This completes the proof of Lemma IV.1.

Proof of Theorem IV.2: Denote $\lambda/p = 2r$. From the optimization criteria (4.8), for any γ^*, we have

$$\frac{1}{n}\|\boldsymbol{Y} - \hat{X}(\hat{\gamma^*}, \boldsymbol{t})\|^2 + 2r\|\hat{\gamma}\|_1 \leq \frac{1}{n}\|\boldsymbol{Y} - \hat{X}(\gamma^*, \boldsymbol{t})\|^2 + 2r\|\gamma\|_1$$

Write $Y_i = \hat{X}(\gamma_0^*, t_i) + \zeta_i = \hat{X}(\gamma_0^*, t_i) + \varepsilon_i + e_i$ where

$$|e_i| = \left|\{X(\gamma_0^*, t_i) - \hat{X}(\gamma_0^*, t_i)\} + \{X(\theta_0, X_0[0], t_i) - X(\gamma_0^*, t_i)\}\right| \leq c_2 r_q + c_3\omega_p = \omega_{p,q}.$$

from Lemma IV.1 and Assumption C3.

Using the expression of $Y_i = \hat{X}(\gamma_0^*, t_i) + \zeta_i$ and applying Assumption C3, we have

$$\begin{aligned}\frac{1}{n}\|\hat{X}(\gamma_0^*, \boldsymbol{t}) - \hat{X}(\hat{\gamma}^*, \boldsymbol{t})\|^2 + 2r\|\hat{\gamma}\|_1 &\leq \frac{1}{n}\|\hat{X}(\gamma_0^*, \boldsymbol{t}) - \hat{X}(\gamma^*, \boldsymbol{t})\|^2 + 2r\|\gamma\|_1 \\ &\quad + \frac{2}{n}\sum_{i=1}^{n}\zeta_i\{\hat{X}(\hat{\gamma}^*, t_i) - \hat{X}(\gamma^*, t_i)\} \\ &\leq \frac{1}{n}\|\hat{X}(\gamma_0^*, \boldsymbol{t}) - \hat{X}(\gamma^*, \boldsymbol{t})\|^2 + 2r\|\gamma\|_1 \\ &\quad + \frac{2}{n}\sum_{k=1}^{p+1}\sum_{i=1}^{n}\mathcal{C}_k^U(t_i)|\zeta_i||\hat{\gamma}_k^* - \gamma_k^*|.\end{aligned}$$

Define the random variable $\tilde{V}_k = n^{-1}\sum_{i=1}^n \mathcal{C}_k^U(t_i)\zeta_i$ and $V_k = n^{-1}\sum_{i=1}^n \mathcal{C}_k^U(t_i)\varepsilon_i$, $1 \leq k \leq p+1$, and the event

$$\mathcal{A} = \bigcap_{k=2}^{p}\{|V_k| \leq r/2 - K_2\omega_{p,q}\}\bigcap\{|V_1| \leq 2r - K_2\omega_{p,q}\},$$

where $K_2 = \bigvee_{k=1}^{p+1}(|\mathcal{M}_{\cdot k}^U|_1/n)$. On the event $\mathcal{A}$ we have $|\tilde{V}_1| \leq 2r$ and $|\tilde{V}_k| \leq r/2$ for $2 \leq k \leq p+1$. Using a bound on the tails of normal distribution, As in Bickel, et al.

(2009), we have that the probability of the complementary event $\mathcal{A}^c$ satisfies

$$\begin{aligned} P(\mathcal{A}^c) &\leq \sum_{k=2}^{p+1} P\{\sqrt{n}|V_k| > \sqrt{n}(r/2 - K_2\omega_{p,q})\} + P\{\sqrt{n}|V_1| > \sqrt{n}(2r - \omega_{p,q})\} \\ &\leq \sum_{k=2}^{p+1} \exp\left\{-\frac{n(r - 2K_2\omega_{p,q})^2}{8\sigma_\varepsilon^2\|\mathcal{M}^U_{\cdot k}\|_n^2}\right\} + \exp\left\{-\frac{n(2r - K_2\omega_{p,q})^2}{2\sigma_\varepsilon^2\|\mathcal{M}^U_{\cdot 1}\|_n^2}\right\} \\ &\leq (p+1)\exp\left\{-\frac{n(r - 2K_2\omega_{p,q})^2}{8\sigma_\varepsilon^2 K_1^2}\right\} = (p+1)^{1-\frac{a^2}{8}}, \end{aligned}$$

where $K_1 = \bigvee_{k=2}^{p+1} \|\mathcal{M}^U_{\cdot k}\|_n \bigvee (4^{-1}\|\mathcal{M}^U_{\cdot 1}\|_n)$ and $r = a\sigma_\varepsilon K_1\sqrt{\log(p+1)/n} + 2K_2\omega_{p,q}$. Then on the event $\mathcal{A}$ we have

$$\frac{1}{n}\|\hat{X}(\gamma_0^*, \boldsymbol{t}) - \hat{X}(\hat{\gamma}^*, \boldsymbol{t})\|^2 \leq \frac{1}{n}\|\hat{X}(\gamma_0^*, \boldsymbol{t}) - \hat{X}(\gamma^*, \boldsymbol{t})\|^2 + 4r|\hat{X}[0] - X[0]| + r\|\hat{\gamma} - \gamma\|_1 + 2r\|\gamma\|_1 - 2r\|\hat{\gamma}\|_1.$$

Denote $\delta = \gamma_0^* - \hat{\gamma}^*$. Applying Assumption C3 and adding the term $r\|\hat{\gamma} - \gamma\|_1$ to both sides of the inequality above yields, on $\mathcal{A}$,

$$\begin{aligned} \frac{1}{n}|\delta|^T\mathcal{M}^{LT}\mathcal{M}^L|\delta| + r|\hat{\gamma} - \gamma|_1 &\leq \frac{1}{n}\|\hat{X}(\gamma_0^*, t) - \hat{X}(\gamma^*, t)\|^2 + 4r|\hat{X}[0] - X[0]| + 4r\sum_{k\in J(\gamma)}|\hat{\gamma}_k - \gamma_k| \\ &\leq \frac{1}{n}\|\hat{X}(\gamma_0^*, t) - \hat{X}(\gamma^*, t)\|^2 + \\ &\quad 4r\sqrt{S(\gamma) + 1}\sqrt{|\hat{X}[0] - X[0]|^2 + \sum_{k\in J(\gamma)}|\hat{\gamma}_k - \gamma_k|^2} \end{aligned}$$

where $S(\gamma) = |J(\gamma)|$. Taking $\gamma^* = \gamma_0^*$ and denoting $J_0 = J(\gamma_0)$, we have $\|\delta_{J_{\mathcal{F}}}\|_1 \leq 4\|\delta^*_{J_0}\|_1$ where $\delta^*_{J_0}$ is defined following the rule of notation given right before Assumption C4. Under Assumption C4 we have

$$\kappa^2\|\delta^*_{J_0}\|^2 \leq 4r\sqrt{S_p + 1}\|\delta^*_{J_0}\|,$$

Therefore,

$$|\hat{X}[0] - X_0[0]| \leq \|\delta^*_{J_0}\| \leq \frac{4r\sqrt{S_p + 1}}{\kappa^2},$$

and

$$\|\hat{\gamma}_0 - \gamma_0\| \leq \|\hat{\gamma}_0 - \gamma_0\|_1 = \|\delta_{J_{\mathcal{F}}}\|_1 \leq 4\sqrt{S_p + 1}\|\delta^*_{J_0}\| \leq \frac{16r(S_p + 1)}{\kappa^2}.$$

Since $\hat{\theta}(\tau) = \psi(\tau)^T \hat{\eta} = \psi(\tau)^T A^{-1}\hat{\gamma}$ and $\theta_0(\tau) = \psi(\tau)^T A^{-1}\gamma_0 + e_p(\tau)$, then we have

$$\begin{aligned} |\hat{\theta}_0(\tau) - \theta_0(\tau)| &\leq |\psi(\tau)^T A^{-1}\hat{\gamma}_0 - \psi(\tau)^T A^{-1}\gamma_0| + |e_p(\tau)| \\ &\leq \|\psi(\tau)^T A^{-1}\| \cdot \|\hat{\gamma}_0 - \gamma_0\| + |e_p(\tau)| \\ &\leq \frac{16\alpha_{n,p}(\tau) r(S_p + 1)}{\kappa^2} + \omega_p. \end{aligned}$$

Furthermore, if $\zeta_i \sim N(0, \sigma_\varepsilon^2)$, simply consider the event

$$\mathcal{A}' = \bigcap_{k=2}^{p} \{|\tilde{V}_k| \leq r/2\} \bigcap \{|\tilde{V}_1| \leq 2r\}.$$

If $r = a\sigma_\varepsilon K_1\sqrt{\log(p+1)/n}$, we have $P(\mathcal{A}'^c) \leq (p+1)^{1-a^2/8}$. Then the rest of the proof will follow the same structure as above. This completes the proof of Theorem IV.2.

Proof of Corollary IV.3: Replacing S_p with $S < \infty$ by (C5), with probability converging to one as $p \to \infty$, we have

$$\begin{aligned} |\hat{\theta}(\tau) - \theta^*(\tau)| &\leq \frac{(p+1)^{b_\tau}\sqrt{\log(p+1)}}{\sqrt{n}} \frac{16a\sigma_\varepsilon K_1(S+1)}{\kappa^2}(p+1)^{-b_\tau}\alpha_{n,p}(\tau) \\ &\quad + \frac{\omega_p(p+1)^\nu}{(p+1)^{\nu - b_\tau}}\left\{(p+1)^{-b_\tau} + \frac{32(p+1)^{-b_\tau}\alpha_{n,p}(\tau)K_2(S+1)}{\kappa^2}\frac{(p+1)^\nu \omega_{p,q}}{(p+1)^\nu \omega_p}\right\} \\ &= \frac{\sqrt{\log n}}{n^{1/2 - b_\tau/(2\nu)}}K \end{aligned}$$

where K equals

$$\begin{aligned} &\left(\frac{p+1}{n^{1/2\nu}}\right)^{b_\tau}\sqrt{\frac{\log(p+1)}{\log n}}\frac{16a\sigma_\varepsilon K_1(S+1)}{\kappa^2}(p+1)^{-b_\tau}\alpha_{n,p}(\tau) \\ &+ \left(\frac{p+1}{n^{1/2\nu}}\right)^{b_\tau - \nu}\frac{\omega_p(p+1)^\nu}{\sqrt{\log n}}\left\{(p+1)^{-b_\tau} + \frac{32(p+1)^{-b_\tau}\alpha_{n,p}(\tau)K_2(S+1)}{\kappa^2}\frac{(p+1)^\nu \omega_{p,q}}{(p+1)^\nu \omega_p}\right\} \end{aligned}$$

Because of $p = O(n^{1/2\nu})$ we have $(p+1)/n^{1/2\nu}$ and $\log(p+1)/\log n$ bounded. $\omega_p(p+1)^\nu$ and $\omega_{p,q}(p+1)^\nu$ are bounded by (C6) and $(p+1)^{-b_\tau}\alpha_{n,p}(\tau)$ is bounded by (C7). Also κ is bounded away from zero from (C9). Hence we have $|\hat{\theta}(\tau) - \theta^*(\tau)| = O(\frac{\sqrt{\log n}}{n^{1/2-b_\tau/(2\nu)}})$. With the addition of the assumption (C8) through exactly the same argument one can prove $\sup_\tau |\hat{\theta}(\tau) - \theta^*(\tau)| = O(\frac{\sqrt{\log n}}{n^{1/2-\mu/(2\nu)}})$. For the initial values, one can easily have $|\hat{X}[0] - X_0[0]| = O(\sqrt{\frac{\log n}{n}})$.

If $\zeta_i \sim N(0, \sigma_\zeta^2)$, we have

$$|\hat{\theta}(\tau) - \theta^*(\tau)| \leq \frac{\sqrt{\log n}}{n^{\nu/(2\nu+2b_\tau)}} K$$

where K equals

$$\left(\frac{p+1}{n^{1/(2\nu+2b_\tau)}}\right)^{b_\tau} \sqrt{\frac{\log(p+1)}{\log n}} \frac{16a\sigma_\zeta K_1(S+1)}{\kappa^2}(p+1)^{-b_\tau}\alpha_{n,p}(\tau) + \left(\frac{p+1}{n^{1/(2\nu+2b_\tau)}}\right)^{-\nu} \frac{\omega_p(p+1)^\nu}{\sqrt{\log n}}$$

Following same arguments, we have $|\hat{\theta}(\tau) - \theta^*(\tau)| = O(\frac{\sqrt{\log n}}{n^{\nu/(2\nu+2b_\tau)}})$. In the same way we have $\sup_\tau |\hat{\theta}(\tau) - \theta^*(\tau)| = O(\frac{\sqrt{\log n}}{n^{\nu/(2\nu+2\mu)}})$.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alterman, R., and E. Stanley (1994), Colony stimulating factor-1 expression in human glioma, *Molecular and Chemical Neuropathology*, *21*, 177–188.

Bard, Y. (1974), Nonlinear Parameter Estimation, *London: Academic.*

Bauch, C. T., and D. J. D. Earn (2003), Transients and attractors in epidemics, *Proceedings of Royal Society B*, *270*, 1573–1578.

Bickel, P., Y. Ritov, and A. Tsybakov (2009), Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics*, *37*, 1705–1732.

Breiman, L. (1995), Better Subset Regression Using the Nonnegative Garrote , *Technometrics*, *37*, 373–384.

Brunel, N.-B. (2008), Parameter estimation of ODEs via nonparametric estimators, *Electron. J. Stat.*, *2*, 242–1267.

Bulter, G. (1974), Principles of Ecotoxicology, *International Council of Scientific Unions, Scientific Committee on Problems of the Environment. John Wiley and Sons, Brisbane, Chichester, New York, Singapore, Toronto*, *SCOPE 12*, 372.

Candes, E., and T. Tao (2005), The Dantzig selector: statistical estimation when p is much larger than n, *Annals of Statistics*, *35*, 23132351.

Cao, J., J. Huang, and H. Wu (2011), Penalized Nonlinear Least Squares Estimation of Time-Varying Parameters in Ordinary Differential Equations, *Journal of Computational and Graphical Statistics*, *ahead of print.*

Chen, J., and H. Wu (2008), Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics, *Journal of the American Statistical Association*, *103*, 369–384.

Dickson, K., A. W. Maki, and J. J. Cairns (1982), Modeling the Fate of Chemical in the Aquatic Environment, *Ann Arbor Science Publishers, Ann Arbor, Michigan.*

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004), Least angle regression, *Annals of Statistics*, *32*, 407–499.

Fan, J., and R. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, *96*, 1348–1360.

Fan, J., and J. Lv (2008), Sure independence screening for ultra-high dimensional feature space, *Journal of the Royal Statistical Society, Series B*, *70*, 849–911.

Fan, J., and H. Peng (2004), Nonconcave penalized likelihood with a diverging number of parameters, *Annals of Statistics*, *32*, 928–961.

FitzHugh, R. (1961), Impulses and Physiological States in Models of Nerve Membrane , *Biophysical Journal*, *1*, 445–466.

Friedman, J., T. Hastie, H. Hoefling, and R. Tibshirani (2007), Pathwise coordinate optimization, *Annals of Applied Statistics*, *1*, 302–332.

Fu, W. (1998), Penalized regression: the bridge versus the Lasso, *Journal of Computational and Graphical Statistics*, *7*, 397–416.

Hall, C. A. (1968), On error bounds for spline interpolation, *J. Approx. Theory*, *1*, 209–218.

Hemker, P. (1972), Numerical Methods for Differential Equations in System Simulation and in Parameter Estimation, *Analysis and Simulation of Biochemical Systems, eds. H.C. Hemker adn B. Hess, Amsterdam and London: North-Holland/New York: American Elsevier*, pp. 59–80.

Hodgkin, A., and A. Huxley (1952), A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve, *Journal of Physiology*, *133*, 444–479.

Hooker, G., S. P. Ellner, L. Roditi, and D. J. D. Earn (2011), Parameterizing state - space models for infectious disease dynamics by generalized profiling: measles in Ontario, *Journal of the Royal Society Interface*, *8*, 961–974.

Horvath, S., et al. (2006), Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target, *Proceedings of National Academy of Sciences*, *103*, 17,402–17,407.

Huang, J., S. Ma, and C.-H. Zhang (2008), Adaptive lasso for sparse high-dimensional regression models, *Statistica Sinica*, *18*, 1603–1618.

Huang, J., X. Ma, S., H., and C.-H. Zhang (2009), A group bridge approach for variable selection, *Biometrika*, *96*, 339–355.

Hutzinger, O. (1985), The Natural Environment and the Biogeochemical Cycles, *The Handbook of Environmental Chemistry, Part D: Springer-Verlag, Birlin, New York, Tokyo*, *1*.

James, G., J. Wang, and J. Zhu (2009), Functional Linear Regression That's Interpretable, *Annals of Statistics*, *37*, 2083–2108.

Kuan, C., K. Wakiya, J. Dowell, J. Herndon, D. Reardon, M. Graner, G. Riggins, C. Wikstrand, and D. Bigner (2006), Glycoprotein nonmetastatic melanoma protein B, a potential molecular therapeutic target in patients with glioblastoma multiforme, *Clin. Cancer Res.*, *12*, 1970–1982.

Lam, C., and J. Fan (2009), Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation, *Annals of Statistics*, *37*, 4254–4278.

Le Mercier, M., F. Lefranc, T. Mijatovic, O. Debeir, B. Haibe-Kains, G. Bontempi, C. Decaestecker, R. Kiss, and V. Mathieu (2008), Evidence of galectin-1 involvement in glioma chemoresistanc, *Toxicol. Appl. Pharmacol.*, *229*, 172–183.

Li, Y., J. Zhu, and N. Wang (2011), Parameter Estimation for Ordinary Differential Equations: An Alternative View on Penalty, *manuscript.*

Liang, H., and H. Wu (2008), Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models , *Journal of the American Statistical Association*, *103*, 1570–1583.

Lin, Y., and H. Zhang (2006), Component Selection and Smoothing in Multivariate Nonparametric Regression , *Annals of Statistics*, *34*, 2272–2297.

Lotka, A. (1910), Contribution to the Theory of Periodic Reaction, *J. Phys. Chem.*, *14*, 271–274.

Meier, L., S. van de Geer, and P. Buhlmann (209), High-Dimensional Additive Modeling, *Annals of Statistics*, *37*, 3779–3821.

Meinshausen, N., and P. Buhlmann (2006), High-dimensional graphs and variable selection with the Lasso, *Annals of Statistics*, *34*, 1436–1462.

Meinshausen, N., and B. Yu (2009), Lasso-type recovery of sparse representations for high-dimensional data, *Annals of Statistics*, *37*, 246–270.

Nagumo, J., S. Arimoto, and S. Yoshizama (1962), An Active Pulse Transmission Line Simulating a Nerve Axon, *Proceeding of the IRE*, *50*, 2061–2070.

Nardi, Y., and A. Rinaldo (2008), On the asymptotic properties of the group Lasso estimator for linear models, *Electronic Journal of Statistics*, *2*, 605–633.

Neely, W. (1980), Chemical in the Environment: Distribution-Transport-Fate-Analysis, *Marcel Dekker,New York.*

Odum, E. (1953), Fundamentals of Ecology, *W. B. Saunders, Philadelphia.*

Qi, X., and H. Zhao (2010), Asymptotic Efficiency and Finite-sample Properties of the Generalized Profiling Estimation of Parameters in Ordinary Differential Equations, *Annals of Statistics*, *38*, 435–481.

Ramsay, J., G. Hooker, D. Campbell, and J. Cao (2007), Parameter Estimation for Differential Equations: A Generalized Smoothing Approach , *Journal of the Royal Statistical Society, Ser. B*, *69*, 741–796.

Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009), Sparse Additive Models, *Journal of the Royal Statistical Society, Series B*, *71*, 1009–1030.

Tibshirani, R. (1996), Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tse, K., et al. (2006), CR011, a fully human monoclonal antibody-auristatin E conjugate, for the treatment of melanoma, *Clin. Cancer Res.*, *12*, 1373–1382.

Varah, J. (1982), A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations, *SIAM Journal on Scientific Computing*, *3*, 131–141.

Volterra, V. (1926), Variazioni e fluttuazioni del numero d'individui in specie animali conviventi, *Mem. Acad. Lincei Roma*, *2*, 31–113.

Wang, S., B. Nan, N. Zhou, and J. Zhu (2009), Hierarchically penalized Cox regression with grouped variables, *Biometrika*, *96*, 307–322.

Yang, Y. (2005), Can the strengths of AIC and BIC be shared?, *Biometrika*, *92*, 937–950.

Yuan, M., and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, *68*, 4967.

Zhang, C.-H., and J. Huang (2008), The sparsity and bias of the Lasso selection in high-dimensional linear regression, *Annals of Statistics*, *36*, 1567–1594.

Zhao, P., and B. Yu (2006), On the model selection consistency of LASSO, *Journal of Machine Learning Research*, *7*, 2541–2563.

Zhao, P., G. Rocha, and B. Yu (2009), The composite absolute penalties family for grouped and hierarchical variable selection, *Annals of Statistics*, *37*, 3468–3497.

Zhou, N., and J. Zhu (2010), Group variable selection via a hierarchical Lasso and its oracle property, *Statistics and Its Interface*, *3*, 557–574.

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2007), On the degrees of freedom of the Lasso, *Annals of Statistics*, *35*, 2173–2192.