Generalized Functional Extended Redundancy Analysis

Heungsun Hwang[1], Hye Won Suk[1]. Yoshio Takane[1], Jang-Han Lee[2], Jooseop Lim[3]

1. McGill University, Montreal, Quebec, Canada

2. Chung-Ang University, Seoul, Korea

3. Concordia University, Montreal, Quebec, Canada

May 30, 2013

The authors thank the Associate Editor and three anonymous reviewers for their constructive comments. The work reported in the paper was supported by the Natural Sciences and Engineering Research Council of Canada (No. 10630) to the third author, by the International Cooperation Program of the National Research Foundation of Korea (2012-K2A1A2033137) to the fourth author, and by the Social Science and Humanities Research Council of Canada and Fonds de Recherche sur la Société et la Culture to the fifth author. Requests for reprints should be sent to: Heungsun Hwang, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada. Email: heungsun.hwang@mcgill.ca.

Accepted for publication in Psychometrika

Abstract

Functional extended redundancy analysis (FERA) was recently developed to integrate data reduction into functional linear models. This technique extracts a component from each of multiple sets of predictor data in such a way that the component accounts for the maximum variance of response data. Moreover, it permits predictor and/or response data to be functional. FERA can be of use in describing overall characteristics of each set of predictor data and in summarizing the relationships between predictor and response data. In this paper, we extend FERA into the framework of generalized linear models (GLM), so that it can deal with response data generated from a variety of distributions. Specifically, the proposed method reduces each set of predictor functions to a component and uses the component for explaining exponential-family responses. As in GLM, we specify the random, systematic, and link function parts of the proposed method. We develop an iterative algorithm to maximize a penalized log-likelihood criterion that is derived in combination with a basis function expansion approach. We conduct two simulation studies to investigate the performance of the proposed method based on synthetic data. In addition, we apply the proposed method to two examples to demonstrate its empirical usefulness.

Keywords: functional data analysis, functional extended redundancy analysis, generalized linear models, data reduction, exponential family responses, penalized likelihood.

1. Introduction

Owing to the invention of novel, sophisticated measurement tools such as eye-tracking devices and functional neuroimaging modalities (e.g., positron emission tomography, functional magnetic resonance imaging, electroencephalography, etc.), researchers in psychology and various fields have increasingly been collecting data in the form of curves, surfaces, or images that vary continuously over time, space, and other continuum. Data of this form are called functional data. There has been a continuing and growing need to analyze such data effectively and gain insight into the data. Functional data analysis (FDA), the term coined by Ramsay and Dalzell (1991), represents a rapidly emerging branch of statistics, which aims to develop and apply statistical techniques for the analysis of functional data.

There is a vast and ever-expanding literature on FDA. A full-range overview is available in Ramsay and Silverman (1997, 2005) and Ferraty and Vieu (2006). The former books provide far-reaching explanations as to various aspects of functional data and functional counterparts of multivariate statistical techniques, while the latter focuses on complementary and extensive treatments on nonparametric techniques for functional data. Past special issues of statistics journals, such as *Computational Statistics*, *Computational Statistics and Data Analysis*, and *Statistica Sinica*, also contain review articles on FDA (e.g., Davidian, Lin, & Wang, 2004; González-Manteiga & Vieu, 2007; Rice, 2004; Valderrama, 2007). A description of more recent developments in FDA can be found in Ferraty (2011), Ferraty, Laksaci, Tadj, and Vieu (2011), Ferraty, Mas and Vieu (2007), Ferraty and Romain (2011), Ferraty, Van Keilegom, & Vieu (2012), and Lian (2011). Moreover, Ramsay and Silverman (2002) showcases a number of enlightening applications across diverse fields. Computational software and support for FDA is available in MATLAB, R, and S-PLUS from Dr. Ramsay's Functional Data Analysis website

(http://www.psych.mcgill.ca/misc/fda/) (also see Ramsay, Hooker & Graves, 2009). Computational software for nonparametric FDA is available from the Nonparametric Functional Data Analysis website (http://www.math.univ-toulouse.fr/staph/npfda/) developed under a research group known as STAPH (Groupe de Travail en Statistique Fonctionnelle et Opeatorielle) and from a group of Spanish researchers at the University of Santiago de Compostela (e.g., Febrero-Bande & Oviedo de la Fuente, 2012).

Functional extended redundancy analysis (FERA) (Hwang, Suk, Lee, Moskowitz, & Lim, 2012) is a recent, technical development in FDA. As the name suggests, this technique is a functional version of extended redundancy analysis (ERA) (Takane & Hwang, 2005). ERA aims to combine data reduction with linear regression simultaneously in the sense that a weighed composite or component is obtained from each of multiple sets of predictor variables such that it accounts for the variance of response variables as much as possible. FERA has the same objective as ERA, but it also permits predictor and/or response data to be functional. Thus, FERA can reduce each set of predictor functional data into a single component and uses the component for explaining response data. By extracting a component from a set of (often high-dimensional) predictor data, FERA aids greatly in summarizing the main characteristics of each set of predictor data and in interpreting the relationships between predictor and response data concisely.

In this paper, we extend FERA into the framework of generalized linear models (GLM) (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972) so as to deal with response data arising from a variety of exponential-family distributions. We call the proposed method generalized functional extended redundancy analysis (GFERA). Technically, the proposed method represents a parametric generalization of FERA, which aims to extract a component from

each of multiple sets of predictor functions in such a way that the component influences exponential-family responses. This method results in maximum-likelihood parameter estimates and their asymptotic standard errors.

The paper is organized as follows. We begin by providing a brief review of GLM in Section 2 and of ERA and FERA in Section 3 in order to help understand their connections with the proposed method. In Section 4, we present the technical details of the proposed method. We provide a specification on the relationship between multiple sets of predictor functions and an exponential-family response variable. We also develop an iterative algorithm to maximize a penalized log-likelihood criterion for parameter estimation. In Section 5, we conduct two simulation studies to evaluate the performance of the proposed method based on synthetic data. In Section 6, we apply the proposed method to two real datasets to illustrate its empirical feasibility. One dataset involves binary responses, whereas the other continuous, normal responses. The final section summarizes the previous sections and discusses prospective extensions of the proposed method.

2. Generalized Linear Models

We need to specify three constituents to formulate a generalized linear model: the distribution of a random response variable (the random part), a linear predictor based on predictor variables (the systematic part), and a link function that connects the random and systematic parts.

Assume that y_i is a single observation on a random response variable arising from a distribution in the exponential family ($i = 1,..., N$). The density function or probability distribution for y_i can be generally expressed as

$$l(y_i;\rho_i,\phi)=\exp\big((y_i\rho_i-\xi(\rho_i))/\zeta(\phi)+\nu(y_i,\phi)\big) \tag{1}$$

for some known functions $\xi(\cdot)$, $\zeta(\cdot)$, and $\nu(\cdot)$. If the dispersion parameter ϕ is known, (1) falls into the exponential family with canonical parameter ρ_i. In (1), y_i is independently distributed with mean μ_i. The dispersion parameter is assumed to be constant over observations. McCullagh and Nelder (1989, p. 30) supply a summary of main characteristics of common univariate distributions in the exponential family.

Let x_{ip} denote the pth predictor variable value for the ith observation ($p=1,...,P$). Let β_0 denote the intercept and β_p denote a regression coefficient for the pth predictor. Let η_i and $g(\cdot)$ denote a linear predictor and a link function, respectively. In GLM, the relationship between a linear predictor and a link function can be specified as

$$\eta_i=\beta_0+\sum_{p=1}^{P}x_{ip}\beta_p=\mathbf{x}_i\boldsymbol{\beta}=g(\mu_i), \tag{2}$$

where $\mathbf{x}_i=[1,x_{i1},...,x_{iP}]$ and $\boldsymbol{\beta}=[\beta_0,\beta_1,...,\beta_P]'$. In particular, when $\rho_i=\eta_i$, we have the canonical link; for example, the identity, logit, log, inverse, and squared inverse functions are the canonical links for the normal, binomial, Poisson, gamma, and inverse Gaussian distributions, respectively (see McCullagh & Nelder, 1989, p. 30).

From (1), the log-likelihood function for N observations is generally given as

$$\varphi_1=\sum_{i=1}^{N}\log l(y_i;\rho_i,\phi). \tag{3}$$

The maximum-likelihood estimates of parameters are typically obtained by using iteratively reweighted least squares (IRLS) (see, e.g., Green, 1984) based on Newton-Raphson or Fisher scoring. Let $\mathbf{u}(\boldsymbol{\beta})$ denote the score vector and $\mathbf{H}(\boldsymbol{\beta})$ denote the Hessian matrix or its expected value. The Newton-Raphson or Fisher-scoring algorithm for maximizing (3) is

$$\boldsymbol{\beta}^{(h+1)} = \boldsymbol{\beta}^{(h)} - \mathbf{H}(\boldsymbol{\beta}^{(h)})^{-1}\mathbf{u}(\boldsymbol{\beta}^{(h)}). \tag{4}$$

This can be written in IRLS form as

$$\boldsymbol{\beta}^{(h+1)} = (\mathbf{X}'\mathbf{V}^{(h)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{(h)}\mathbf{t}^{(h)}, \tag{5}$$

where $\mathbf{X} = \left[\mathbf{x}_1,...,\mathbf{x}_N\right]'$, $\mathbf{V}^{(h)}$ is an N by N diagonal matrix of weights with elements $v_i^{(h)} = \dfrac{(\partial\mu_i^{(h)} / \partial\eta_i^{(h)})^2}{\tau_i^{(h)}}$, where $\tau_i^{(h)}$ is the variance function evaluated at $\mu_i^{(h)}$, and $\mathbf{t}^{(h)}$ is an N by 1 vector of the so-called adjusted response variable with elements $t_i^{(h)} = \eta_i^{(h)} + (y_i - \mu_i^{(h)}) / v_i^{(h)}$ (refer to McCullagh & Nelder, 1989, Chapter 2). Thus, obtaining $\boldsymbol{\beta}^{(h+1)}$ in (5) is equivalent to minimizing the following generalized least-squares criterion

$$\varphi_2 = (\mathbf{t}^{(h)} - \mathbf{X}\boldsymbol{\beta}^{(h+1)})'\mathbf{V}^{(h)}(\mathbf{t}^{(h)} - \mathbf{X}\boldsymbol{\beta}^{(h+1)}), \tag{6}$$

with respect to $\boldsymbol{\beta}^{(h+1)}$ (see, e.g., Yee & Hastie, 2003). This iterative procedure continues until no substantial differences between $\boldsymbol{\beta}^{(h+1)}$ and $\boldsymbol{\beta}^{(h)}$ occur.

Once the maximum-likelihood estimates, denoted by $\hat{\boldsymbol{\beta}}$, are obtained, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ can be obtained by

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}\right)^{-1}, \tag{7}$$

where $\hat{\mathbf{V}}$ is the matrix of weight estimates obtained at convergence.

3. Extended Redundancy Analysis and Functional Extended Redundancy Analysis

In this section, we briefly explain ERA and FERA to facilitate an understanding of the derivation of the proposed method. The original ERA can accommodate scalar and multivariate response data and FERA can deal with scalar, multivariate, or functional response data. However,

GLM typically considers a scalar response variable. Thus, we focus on the formulation of ERA and FERA models that deal with a scalar response variable with multiple sets of predictor data.

3.1. Extended Redundancy Analysis

Let x_{ikp} denote the pth variable value in the kth set of predictor variables on the ith observation ($k = 1,...,K$; $p = 1,...,P_k$). Let π_{kp} denote a component weight value assigned to x_{ikp}. Let a_k denote the kth component loading connecting the kth component to the response variable y_i. Let e_i denote the ith residual value for y_i.

The ERA model for a scalar response variable can be written as

$$\begin{aligned} y_i &= \sum_{k=1}^{K} \left[\sum_{p=1}^{P_k} x_{ikp} \pi_{kp} \right] a_k + e_i \\ &= \sum_{k=1}^{K} f_{ik} a_k + e_i , \end{aligned} \tag{8}$$

where $f_{ik} = \left[\sum_{p=1}^{P_k} x_{ikp} \pi_{kp} \right]$ indicates the ith scalar component score of the kth set of predictor variables. As shown in (8), the ERA model typically assumes that a single component or weighted composite of each set of predictor variables influences the response variable. These influences are represented by component loadings.

Figure 1 displays an example of an ERA model. This model involves three sets of predictor variables (K = 3), each of which consists of two variables (P_k = 2). Each component is constructed by adding component weights to the corresponding set of predictor variables, and influences a response variable. As illustrated in the figure, the ERA model can be viewed as a special type of structural equation model, in which all exogenous latent variables are equivalent

to components of predictor variables and the endogenous variable is always observed (Takane & Hwang, 2005).

Insert Figure 1 about here

To estimate parameters, the following least squares criterion is minimized

$$\varphi_3 = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} \left[\sum_{p=1}^{P_k} x_{ikp} \pi_{kp} \right] a_k \right)^2 = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} f_{ik} a_k \right)^2, \quad (9)$$

with respect to π_{kp} and a_k, subject to $\sum_{i=1}^{N} f_{ik}^{\ 2} = N$, which is a conventional scaling constraint on each set of component scores. Takane and Hwang (2005) proposed an alternating least-squares algorithm to minimize (9). This algorithm alternates two main steps until convergence. In the first step, with a_k fixed, π_{kp} is updated in a least squares sense. In the second, with π_{kp} fixed, a_k is updated in a least squares sense.

3.2. Functional Extended Redundancy Analysis

Let z_{ik} denote the ith function in the kth set of predictor functions ($k = 1,\ldots, K$) with values $z_{ik}(s_k)$ for any argument s_k ($s_k \in S_k$). Let w_k denote the kth component weight function associated with the kth set of predictor functions with values $w_k(s_k)$.

We posit that the component of each set of predictor functions influences the scalar response variable. The FERA model for this case can be given as

$$
\begin{aligned}
y_i &= \sum_{k=1}^{K} \left[\int_{S_k} z_{ik}(s_k) w_k(s_k) ds_k \right] a_k + e_i \\
&= \sum_{k=1}^{K} f_{ik} a_k + e_i,
\end{aligned}
\tag{10}
$$

where $f_{ik} = \int_{S_k} z_{ik}(s_k) w_k(s_k) ds_k$ denotes the ith scalar component score of the kth set of predictor functions. In (10), both z_{ik} and w_k are functions, so that the component score f_{ik}, which is the inner product of these functions, is defined as integration over S_k.

To estimate parameters in (10), we minimize the following penalized least-squares criterion

$$
\varphi_4 = \sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{K} \left[\int_{S_k} z_{ik}(s_k) w_k(s_k) ds_k \right] a_k \right)^2 + \lambda \sum_{k=1}^{K} \int_{S_k} [D^2 w_k(s_k)]^2 ds_k, \tag{11}
$$

with respect to w_k and a_k, subject to $\sum_{i=1}^{N} {f_{ik}}^2 = N$. In (11), λ is a non-negative smoothing parameter that controls for the degree of roughness in w_k; $\sum_{k=1}^{K} \int \left[D^2 w_k(s_k) \right]^2 ds_k$ denotes a roughness penalty term that is the sum of the integrated squared second derivative of w_k over K sets; and D^c denotes the derivative of order c.

Criterion (11) is derived by integrating a roughness penalty term into the ordinary least-squares criterion. This roughness penalty term is introduced to take account of the degree of a function's roughness or curvature. A function's curvature at argument r is measured by its second derivative at r, as a straight line has no curvature, producing a zero second derivative. Accordingly, a function's roughness is usually assessed by its integrated squared second derivative (Ramsay & Silverman, 2005, p. 84). In (11), such a roughness penalty term is added for w_k. When w_k is highly variable, its roughness penalty value tends to be large.

The smoothing parameter λ plays a role in balancing out the relative importance of the roughness penalty term in estimation of the weight function. If the smoothing parameter value becomes large, then a greater penalty is imposed on the roughness of the estimated weight function. This leads to keeping the degree of the function' roughness small and to making the function less variable.

An alternating regularized least-squares algorithm (Hwang, 2009) was developed to minimize (11) in combination with a basis function expansion approach to approximating both predictor and weight functions. In the next section, we will discuss a basis function expansion approach for the proposed method in detail.

4. Generalized Functional Extended Redundancy Analysis

We now extend the FERA model (10) into the framework of GLM. This extension conceptually involves two main steps that are to be carried out simultaneously. One step is to reduce a set of intrinsically infinite-dimensional predictor functions to a single variable, called a component, which is the inner product of predictor functions and weight functions. The other is to obtain the component in such a way that it is highly related to a response variable generated from an exponential-family distribution. Computationally, the proposed method begins by uncovering functions by projecting their observations onto a finite number of known functions called basis functions. It subsequently involves optimization of a penalized maximum-likelihood criterion to estimate parameters in a manner similar to GLM.

Let a_0 denote the intercept. We specify the relationship between a linear predictor and a link function as follows.

$$\begin{aligned}\eta_i &= a_0 + \sum_{k=1}^{K}\left[\int z_{ik}(s_k)w_k(s_k)ds_k\right]a_k \\ &= a_0 + \sum_{k=1}^{K} f_{ik}a_k = \mathbf{f}_i\mathbf{a} \\ &= g(\mu_i),\end{aligned} \tag{12}$$

where z_{ik}, w_k, a_k, and f_{ik} are defined in the previous section, $\mathbf{f}_i = [1, f_{i1},..., f_{iK}]$, and $\mathbf{a} = [a_0, a_1,..., a_K]'$. In (12), the interval S_k for an integral is omitted to make the notation concise. This will be the case for all the equations that follow.

To estimate parameters, we seek to maximize a penalized log-likelihood criterion taking the general form as

$$\varphi_5 = \sum_{i=1}^{N}\log l(y_i;\rho_i,\phi) - \frac{1}{2}\lambda\sum_{k=1}^{K}\int[D^2 w_k(s_k)]^2 ds_k. \tag{13}$$

This criterion can be viewed as the L_2-norm penalized log-likelihood because the roughness penalty term can belong to the L_2-norm or quadratic penalty which includes the well-known ridge penalty (Hoerl & Kennard, 1970; see also Le Cessie & Van Houwelingen, 1992; Lee & Silvapulle, 1988).

As discussed in Section 2, maximizing (13) via IRLS is equivalent to minimizing the following penalized generalized least-squares criterion

$$\begin{aligned}\varphi_6 &= \sum_{i=1}^{N} v_i(t_i - a_0 - \sum_{k=1}^{K}\left[\int z_{ik}(s_k)w_k(s_k)ds_k\right]a_k)^2 + \lambda\sum_{k=1}^{K}\int[D^2 w_k(s_k)]^2 ds_k \\ &= (\mathbf{t}-\mathbf{F}\mathbf{a})'\mathbf{V}(\mathbf{t}-\mathbf{F}\mathbf{a}) + \lambda\sum_{k=1}^{K}\int[D^2 w_k(s_k)]^2 ds_k,\end{aligned} \tag{14}$$

with respect to w_k and a_k, subject to $\sum_{i=1}^{N} {f_{ik}}^2 = N$, where $\mathbf{F} = [\mathbf{f}_1,...,\mathbf{f}_N]'$.

In FDA, a function can be represented as a linear combination of known basis functions (Hastie, Tibshirani, & Friedman, 2009; Ramsay & Silverman, 2005, Chapter 3). The central idea

of this basis function expansion approach is that any infinite-dimensional function can be approximated to some arbitrary degree of precision by a linear combination of a finite number of known basis functions. The predictor and weight functions in (12) can be generally expressed as basis function systems as follows.

$$z_{ik}(s_k) = \sum_{m=1}^{M_k} c_{imk}\psi_{mk}(s_k) = \mathbf{c}_{ik}'\boldsymbol{\psi}_k(s_k) \text{ , and } w_k(s_k) = \sum_{m=1}^{M_k} \theta_{mk}\psi_{mk}(s_k) = \boldsymbol{\psi}_k(s_k)'\boldsymbol{\theta}_k, \quad (15)$$

where $\boldsymbol{\psi}_k(s_k)$ is an M_k by 1 vector of basis functions for predictor and weight functions, $\mathbf{c}_{ik}$ and $\boldsymbol{\theta}_k$ are M_k by 1 vectors of coefficients associated with $\boldsymbol{\psi}_k(s_k)$. Let $\mathbf{z}_k(s_k) = \left[z_{1k}(s_k),...,z_{Nk}(s_k)\right]'$. We can then express an N by 1 vector of predictor functions at argument s_k as

$$\mathbf{z}_k(s_k) = \mathbf{C}_k\boldsymbol{\psi}_k(s_k), \quad (16)$$

where $\mathbf{C}_k = \left[\mathbf{c}_{1k},...,\mathbf{c}_{Nk}\right]'$.

A class of basis functions is available for representing functions. For instance, a Fourier series can be used to approximate very stable, periodic functions. B-splines can be chosen for non-periodic functions that may fluctuate locally. Refer to Ramsay and Silverman (2005, Chapter 3) for a description of different basis function systems.

Using the basis function expansion approach, we can express (14) as

$$
\begin{aligned}
\varphi_6 &= \sum_{i=1}^{N} v_i (t_i - a_0 - \sum_{k=1}^{K} \left[\int z_{ik}(s_k) w_k(s_k) ds_k \right] a_k)^2 + \lambda \sum_{k=1}^{K} \int [D^2 w_k(s_k)]^2 ds_k \\
&= \sum_{i=1}^{N} v_i (t_i - a_0 - \sum_{k=1}^{K} \mathbf{c}_{ik}' \left[\int \boldsymbol{\psi}_k(s_k) \boldsymbol{\psi}_k(s_k)' ds_k \right] \boldsymbol{\theta}_k a_k)^2 + \lambda \sum_{k=1}^{K} \int [D^2 \boldsymbol{\psi}_k(s_k)' \boldsymbol{\theta}_k]^2 ds_k \\
&= (\mathbf{t} - a_0 \mathbf{1} - \sum_{k=1}^{K} \mathbf{C}_k \mathbf{J}_k \boldsymbol{\theta}_k a_k)' \mathbf{V} (\mathbf{t} - a_0 - \sum_{k=1}^{K} \mathbf{C}_k \mathbf{J}_k \boldsymbol{\theta}_k a_k) + \lambda \sum_{k=1}^{K} \boldsymbol{\theta}_k' \left[\int D^2 \boldsymbol{\psi}_k(s_k) D^2 \boldsymbol{\psi}_k(s_k)' ds_k \right] \boldsymbol{\theta}_k \\
&= (\mathbf{t} - a_0 \mathbf{1} - \sum_{k=1}^{K} \mathbf{f}_k a_k)' \mathbf{V} (\mathbf{t} - a_0 - \sum_{k=1}^{K} \mathbf{f}_k a_k) + \lambda \sum_{k=1}^{K} \boldsymbol{\theta}_k' \mathbf{R}_k \boldsymbol{\theta}_k \\
&= (\mathbf{t} - [\mathbf{1}, \mathbf{f}_1, ..., \mathbf{f}_K] \mathbf{a})' \mathbf{V} (\mathbf{t} - [\mathbf{1}, \mathbf{f}_1, ..., \mathbf{f}_K] \mathbf{a}) + \lambda \sum_{k=1}^{K} \boldsymbol{\theta}_k' \mathbf{R}_k \boldsymbol{\theta}_k \\
&= (\mathbf{t} - \mathbf{F}\mathbf{a})' \mathbf{V} (\mathbf{t} - \mathbf{F}\mathbf{a}) + \lambda \sum_{k=1}^{K} \boldsymbol{\theta}_k' \mathbf{R}_k \boldsymbol{\theta}_k,
\end{aligned}
\tag{17}
$$

where $\mathbf{1}$ is an N by 1 vector of ones, $\mathbf{C}_k = [\mathbf{c}_{1k}, ..., \mathbf{c}_{Nk}]'$, $\mathbf{J}_k = \int \boldsymbol{\psi}_k(s_k) \boldsymbol{\psi}_k(s_k)' ds_k$, $\mathbf{R}_k = \int D^2 \boldsymbol{\psi}_k(s_k) D^2 \boldsymbol{\psi}_k(s_k)' ds_k$, $\mathbf{f}_k = \mathbf{C}_k \mathbf{J}_k \boldsymbol{\theta}_k$, and $\mathbf{F} = [\mathbf{1}, \mathbf{f}_1, ..., \mathbf{f}_K]$. Thus, maximizing (13) becomes essentially equivalent to solving a regularized linear regression problem through the adoption of the basis function expansion approach. The basis function expansion helps reduce an infinite-dimensional functional problem into a finite-dimensional one which involves a vector of unknown coefficients as in conventional multivariate statistics.

To minimize (17), we may use an iterative algorithm similar to the alternating regularized least-squares algorithm. This algorithm repeats the following steps until no substantial changes in parameter estimates occur.

Step 1: We update $\mathbf{a}$ for fixed $\boldsymbol{\theta}_k$, $\mathbf{V}$, and $\mathbf{t}$. This is equivalent to minimizing

$$
\varphi_7 = (\mathbf{t} - \mathbf{F}\mathbf{a})' \mathbf{V} (\mathbf{t} - \mathbf{F}\mathbf{a}), \tag{18}
$$

with respect to $\mathbf{a}$. Hence, the estimates of $\mathbf{a}$ are obtained by

$$
\hat{\mathbf{a}} = (\mathbf{F}'\mathbf{V}\mathbf{F})^{-1} \mathbf{F}'\mathbf{V}\mathbf{t}. \tag{19}
$$

Step 2: We update $\boldsymbol{\theta}_k$ for fixed $\mathbf{a}$, $\mathbf{V}$, and $\mathbf{t}$. Let $\boldsymbol{\theta} = [\boldsymbol{\theta}_1; ...; \boldsymbol{\theta}_K]$. Criterion (17) can be written as

$$
\begin{aligned}
\varphi_6 &= (\mathbf{t}-\mathbf{Fa})'\mathbf{V}(\mathbf{t}-\mathbf{Fa}) + \lambda\sum_{k=1}^{K}\boldsymbol{\theta}_k'\mathbf{R}_k\boldsymbol{\theta}_k \\
&= (\mathbf{t}-[\mathbf{1},\mathbf{C}_1\mathbf{J}_1,...,\mathbf{C}_K\mathbf{J}_K]\begin{bmatrix}1 & & & \\ & \boldsymbol{\theta}_1 & & \\ & & \mathrm{O} & \\ & & & \boldsymbol{\theta}_K\end{bmatrix}\mathbf{a})'\mathbf{V}(\mathbf{t}-[\mathbf{1},\mathbf{C}_1\mathbf{J}_1,...,\mathbf{C}_K\mathbf{J}_K]\begin{bmatrix}1 & & & \\ & \boldsymbol{\theta}_1 & & \\ & & \mathrm{O} & \\ & & & \boldsymbol{\theta}_K\end{bmatrix}\mathbf{a}) \\
&\quad + \lambda\boldsymbol{\theta}'\begin{bmatrix}\mathbf{R}_1 & & \\ & \mathrm{O} & \\ & & \mathbf{R}_K\end{bmatrix}\boldsymbol{\theta} \\
&= (\mathbf{t}-\mathbf{MQa})'\mathbf{V}(\mathbf{t}-\mathbf{MQa}) + \lambda\boldsymbol{\theta}'\mathbf{G}\boldsymbol{\theta} \\
&= (\mathbf{t}-(\mathbf{a}'\otimes\mathbf{M})\mathrm{vec}(\mathbf{Q}))'\mathbf{V}(\mathbf{t}-(\mathbf{a}'\otimes\mathbf{M})\mathrm{vec}(\mathbf{Q})) + \lambda\boldsymbol{\theta}'\mathbf{G}\boldsymbol{\theta} \\
&= (\mathbf{t}-\boldsymbol{\Sigma}\boldsymbol{\theta})'\mathbf{V}(\mathbf{t}-\boldsymbol{\Sigma}\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}'\mathbf{G}\boldsymbol{\theta},
\end{aligned}
\tag{20}
$$

where $\mathbf{M} = [\mathbf{1},\mathbf{C}_1\mathbf{J}_1,...,\mathbf{C}_K\mathbf{J}_K]$, $\mathbf{Q} = \begin{bmatrix}1 & & & \\ & \boldsymbol{\theta}_1 & & \\ & & \mathrm{O} & \\ & & & \boldsymbol{\theta}_K\end{bmatrix}$, $\mathbf{G} = \begin{bmatrix}\mathbf{R}_1 & & \\ & \mathrm{O} & \\ & & \mathbf{R}_K\end{bmatrix}$, vec(**L**) indicates a supervector formed by stacking all columns of **L** one below another, $\otimes$ indicates the Kronecker product, $\boldsymbol{\theta}$ is equivalent to the vector formed by eliminating such fixed elements as one and zeros from vec(**Q**), $\boldsymbol{\Sigma}$ is the matrix formed by eliminating the columns of $\mathbf{a}'\otimes\mathbf{M}$ corresponding to the fixed elements in vec(**Q**).

Then, the estimates of $\boldsymbol{\theta}$ are obtained by

$$
\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Sigma}'\mathbf{V}\boldsymbol{\Sigma} + \lambda\mathbf{G}\right)^{-1}\boldsymbol{\Sigma}'\mathbf{V}\mathbf{t}. \tag{21}
$$

Subsequently, we obtain $\mathbf{f}_k$ by $\mathbf{f}_k = \mathbf{C}_k\mathbf{J}_k\boldsymbol{\theta}_k$ and standardize it to satisfy $\mathbf{f}_k'\mathbf{f}_k = N$.

Step 3: We update **V** and **t** for fixed **a** and $\boldsymbol{\theta}_k$. As discussed in Section 2, **t** is updated based on $t_i^{(h)} = \eta_i^{(h)} + (y_i - \mu_i^{(h)})/v_i^{(h)}$. The calculation of **V** depends on which distribution is assumed for responses. For example, for a normal distribution, **V** = **I**, and for a binomial distribution, **V** has

elements $v_i = \vartheta_i(1-\vartheta_i)$, where $\vartheta_i = \exp(\eta_i)/(1+\exp(\eta_i))$. Refer to McCullagh and Nelder (1989) for calculation of $\mathbf{V}$ for other exponential-family distributions.

Let $\hat{\boldsymbol{\gamma}} = [\hat{\mathbf{a}}; \hat{\boldsymbol{\theta}}]$ denote the maximum-likelihood parameter estimates at convergence. The asymptotic covariance matrix of $\hat{\boldsymbol{\gamma}}$ can be obtained by computing the negative Hessian matrix evaluated at $\hat{\boldsymbol{\gamma}}$ and inverting it. The negative Hessian matrix is given as

$$-\mathbf{H}(\boldsymbol{\gamma}) = -\frac{\partial^2 \varphi_5}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} = -\begin{bmatrix} \frac{\partial^2 \varphi_5}{\partial \mathbf{a} \partial \mathbf{a}'} & \frac{\partial^2 \varphi_5}{\partial \mathbf{a} \partial \boldsymbol{\theta}'} \\ \frac{\partial^2 \varphi_5}{\partial \boldsymbol{\theta} \partial \mathbf{a}'} & \frac{\partial^2 \varphi_5}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \end{bmatrix}. \tag{22}$$

In (22), $\frac{\partial^2 \varphi_5}{\partial \mathbf{a} \partial \mathbf{a}'}$ and $\frac{\partial^2 \varphi_5}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$ may be obtained by fixing $\boldsymbol{\theta}$ and $\mathbf{a}$, respectively, and are straightforward to calculate. Conversely, $\frac{\partial^2 \varphi_5}{\partial \mathbf{a} \partial \boldsymbol{\theta}'}$ may be computed using the profile likelihoods (Richards, 1961; also see Yee & Hastie, 2003)

$$-\frac{\partial^2 \varphi_5}{\partial \mathbf{a} \partial \boldsymbol{\theta}'} = -\frac{\partial \boldsymbol{\theta}'}{\partial \mathbf{a}}\left(-\frac{\partial^2 \varphi_5}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right). \tag{23}$$

Let $\boldsymbol{\delta}_d$ denote a P+1 by 1 vector of zeros except having one in the dth element (d = 1,…, P+1). Specifically, in the proposed method, $\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{a}}$ is calculated by

$\frac{\partial \boldsymbol{\theta}}{\partial a_d} = \left(\boldsymbol{\Sigma}'\mathbf{V}\boldsymbol{\Sigma} + \lambda\mathbf{G}\right)^{-2}\left[\left(\boldsymbol{\Sigma}'\mathbf{V}\boldsymbol{\Sigma} + \lambda\mathbf{G}\right)\boldsymbol{\Lambda}'\mathbf{V}\mathbf{t} - \boldsymbol{\Lambda}'\mathbf{V}\boldsymbol{\Sigma}\boldsymbol{\Sigma}'\mathbf{V}\mathbf{t}\right]$, where $\boldsymbol{\Lambda}$ is the matrix formed by eliminating the columns of $\boldsymbol{\delta}_d' \otimes \mathbf{M}$ corresponding to the fixed elements in vec($\mathbf{Q}$).

We need to specify the value of λ prior to using the iterative parameter estimation procedure. We may determine the value of λ subjectively (Ramsay & Silverman, 2005, p. 206). We may also use J-fold cross validation (Hastie et al., 2009, p. 214) to choose the value of λ in

an automatic manner. In *J*-fold cross validation, we divide the entire set of data into *J* subsets. We leave one subset as a test set and use the remaining subsets as a training set for estimating parameters. We apply the parameter estimates obtained from the training set to the test set, and calculate its (minus) log-likelihood value. We repeat this procedure *J* times, varying test and training sets systematically. Then, we compute the average of the (minus) log-likelihood values over all *J* test sets.

We may consider a range of alternative values of λ and reiterate the above procedure for each alternative. The value of λ associated with the largest average log-likelihood value (or equivalently, the smallest average minus log-likelihood value) may be chosen as the final one.

A good choice on the number of basis functions (M_k) is important in the estimation of smooth functions. In general, the smaller the number of basis functions, the smoother a function. An excessively large number of basis functions (e.g., $M_k > N$) may be less efficient computationally, whereas a too small number of basis functions may be less satisfactory to capture an important local variation of a function. In the proposed method, as described above, the smoothing parameter (λ) is selected automatically to control for smoothness of parameter functions in combination with a basis function expansion approach. This regularization helps us to choose the number of basis functions in a flexible manner, because similar estimates of parameter functions can be obtained by changing the value of the smoothing parameter, regardless of how many basis functions are used. For example, if a large number of basis functions are used, a smaller value of the smoothing parameter tends to be chosen by cross validation, whereas if a small number of basis functions are used, a larger value of the smoothing parameter tends to be chosen. In the proposed method, it may be convenient computationally to use a relatively small number of basis functions.

5. Simulation Studies

5.1. Simulation Study 1

In the first simulation study, we investigated how well the proposed method recovered weight functions and loadings under two different types of data generated from normal and binomial distributions. For the study, we considered four sample sizes: N = 50, 100, 200, and 500. We considered two sets of predictor functions (K = 2) and $s_k \in [0,100]$. We chose the values of an intercept (a_0) and two loadings (a_1 and a_2) as follows: a_0 = .3, a_1 = .7, and a_2 = -.8. We generated two weight functions based on B-spline basis function expansions. Specifically, we assumed that the two weight functions were represented by 10 B-spline basis functions as follows

$$w_k(s_k) = \sum_{m=1}^{10} \psi_{mk}(s_k)\theta_{mk} = \boldsymbol{\psi}_k(s_k)'\boldsymbol{\theta}_k,$$

where $\boldsymbol{\psi}_k(s_k)$ is a vector of 10 B-spline basis functions over $s_k \in [0,100]$ with equally-spaced knots, and $\boldsymbol{\theta}_k$ is a 10 by 1 vector of basis function coefficients given as:

$$\boldsymbol{\theta}_1 = \left[\, 1,\, 1,\, 1,\, 0,\, 0,\, 0,\, 0,\, 0,\, 0,\, 0 \,\right]'$$
$$\boldsymbol{\theta}_2 = \left[\, 0,\, 0,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1,\, 1,\, 1 \,\right]'.$$

Figure 2 displays the two weight functions generated for the study.

Insert Figure 2 about here

Similarly, the two sets of predictor functions were represented by 10 B-spline basis functions:

$$z_{ik}(s_k) = \sum_{m=1}^{10} c_{imk}\psi_{mk}(s_k) = \mathbf{c}_{ik}'\boldsymbol{\psi}_k(s_k),$$

where $\mathbf{c}_{ik}$ is a 10 by 1 vector of basis function coefficients whose elements were randomly chosen from the standard normal distribution. The predictor functions were obtained once for each sample size, and then considered fixed.

For a normal case, the response variable was generated from $y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \eta_i$ in (12), and σ^2 was fixed to .5, which corresponded approximately to the signal-to-noise ratio of 1/2. For a binomial case, the response variable was generated from $y_i \sim B(1, \mu_i)$, where $\mu_i = \exp(\eta_i)/(1+\exp(\eta_i))$.

We generated 100 samples randomly from each distribution per sample size. We applied five-fold cross validation to select the value of λ for each sample. We considered different values of λ, varying from 0 to 10 by 0.5 in the common logarithmic scale.

To assess the degree of parameter recovery of the loadings and weight functions, we calculated the mean square error (MSE) of each estimate obtained from the proposed method. The MSE is calculated as

$$\text{MSE} = E\left[(\hat{\omega}-\omega)^2\right] = \left(\hat{\omega}-E(\hat{\omega})\right)^2 + \left(E(\hat{\omega})-\omega\right)^2, \quad (24)$$

where ω and $\hat{\omega}$ denote a parameter and its estimate, respectively. In (24), the first and second terms denote the variance and squared bias of the estimate, respectively.

Table 1 provides the MSE, squared biases, and variances of loading estimates obtained from the normal case. The squared biases of the loading estimates decreased with sample size and were quite close to zeros across all sample sizes. The variances of the estimates were also small and approached zeros with sample size. Consequently, the MSE values of the estimates decreased with sample size and came close to zeros. Table 2 shows the MSE, squared biases, and variances of loading estimates obtained from the binomial case. Similarly to the normal case, the squared biases of the loading estimates tended to approach zeros when sample size increased.

The variances of the estimates decreased with sample size. Thus, their MSE values decreased with sample size.

Figures 3, 4, and 5 exhibit the MSE, squared biases, and variances of the two weight functions estimated in the normal case. The squared biases of the estimated weight functions decreased with sample size and were quite small across all sample sizes. Their variances also decreased with sample size. Thus, the MSE values of the estimated weight functions decreased with sample size and appeared small in general. Figures 6, 7, and 8 display the MSE, squared biases, and variances of the two weight functions estimated in the binomial case. As in the normal case, both squared biases and variances of the estimated weight functions decreased with sample size. Thus, their MSE values decreased with sample size. For each sample size, there existed fluctuations in magnitude over the trajectory of each of the three properties in the normal and binomial cases. In particular, there seemed to be more fluctuations in small samples. The patterns of fluctuation were irregular, so that it was difficult to explain them in a clean manner. Nevertheless, the degree of fluctuation appeared to be relatively small. For example, for the normal data, the highest degree of fluctuation in the trajectory of the MSE values (the difference between the largest and smallest MSE values) across sample sizes was approximately 0.2 when $N = 50$, whereas for the binomial data, it was around 0.5 when $N = 50$.

In sum, the simulation study shows that on average, the proposed method was able to recover both loadings and weight functions sufficiently well, when sample size increased.

Insert Figures 3, 4, 5, 6, 7, and 8 about here

5.2. Simulation Study 2

The second study focused on the comparison of parameter recovery between the proposed method and a two-step, sequential approach. In the two-step approach, a functional principal component analysis (e.g., Dauxois, Pousse, & Romain, 1982; Rice & Silverman, 1991; Ramsay & Silverman, 2005, chapter 8) was applied to each set of predictor functions to obtain the first principal component and subsequently, the component was used to predict a response variable generated from normal and binomial distributions. For this study, we used the same synthetic data generated for the first simulation study. This indicates that the solutions obtained from the proposed method remained the same.

Table 1 provides the MSE, squared biases, and variances of loading estimates obtained from the two-step approach in the normal case. The MSE, squared biases, and variances of the intercept estimate (a_0) obtained from the two-step approach were almost identical to those from the proposed method across sample sizes. All property values decreased with sample size. The variances of the other loading estimates (a_1 and a_2) obtained under the two-step approach were slightly smaller than or quite similar to those under the proposed method. They decreased with sample size. On the other hand, the squared biases of the two loading estimates obtained under the two-step approach did not decrease monotonically with sample size. As a result, their MSE values did not decrease monotonically as sample size increased.

Table 2 presents the MSE, squared biases, and variances of loading estimates obtained from the two-step approach in the binomial case. As in the normal case, the MSE, squared biases, and variances of the intercept estimate obtained under the two-step approach were smaller than or similar to those under the proposed method across sample sizes. All of them decreased with sample size. The variances of the other loading estimates obtained under the two-step approach were smaller than those under the proposed method across all sample sizes. They also decreased

with sample size. On the other hand, the squared biases of the same loading estimates were much greater than the counterparts obtained under the proposed method across sample sizes. Moreover, they did not decrease monotonically with sample size. Thus, their MSE values did not decrease with sample size and were greater than those under the proposed method except for $N = 50$. In the two-step approach, a functional principal component analysis was applied to extract components to explain the variance of the predictor functions without taking into account the relationships between the predictor functions and dependent variable. Moreover, a different set of predictor functions was considered for each sample size. Thus, it is hardly expected that the effects of these components on the dependent variable would change systematically with sample size (e.g., the effects continue to be closer to their true parameters when sample size increases). This might explain the somewhat erratic patterns of the MSE values of the loading estimates obtained from the two-step approach.

Figures 9 and 10 display the MSE values of the two weight functions estimated from the two-step approach in the normal and binomial cases, respectively. As stated earlier, in our simulation studies, two predictor functions were generated once and then considered fixed for each sample size. That is, the predictor functions remained the same while generating 100 random samples at each sample size. Thus, when we applied a functional principal component analysis to estimate weight functions for the predictor functions, we ended up obtaining 100 sets of the same weight function estimates for each sample size. This indicates that the weight functions estimated under the two-step approach had no variances, so their MSE values were equal to their squared biases. Accordingly, for the two-step approach, we reported only the MSE values of the estimated weight functions. In general, these MSE values were much greater than those from the proposed method across sample sizes. Moreover, the MSE values under the two-

step approach did not seem to decrease monotonically with sample size. Thus, we can conclude that in both normal and binomial cases, overall, the proposed method yielded estimates that were closer to the true parameter values, as compared to the two-step approach.

Insert Figures 9 and 10 about here

---

6. Empirical Illustrations

6.1. The Lie Detection Data

The first example is part of the data collected from a mock crime experiment, which was designed to investigate the usefulness of various physiological measures in detecting a lie on a crime. In this experiment, 54 participants, who were college students aged between 18 and 29, were divided into two groups, guilty and innocent, according to their own choice. Among these participants, 28 chose to be in the guilty group and 26 in the innocent group. Guilty participants were to commit a mock crime; they were asked to go to a computer room and to take the wallet of a person sitting in a specified seat. Innocent participants were not to commit a mock crime; they were asked to go to the computer room and to send an email to the experimenter, sitting in any seat as they wanted. After coming back to the experimental lab, all participants took a guilty knowledge test, while simultaneously measuring various physiological measures including skin conductance level and pulse. Skin conductance level indicates electrical conductance of the skin that varies with its moisture or sweat level. Pulse indicates heart rate calculated from blood volume pulse, which measures the volume of blood flowing through the extremities. Both skin

conductance level and heart rate are known to be associated with emotions (Kreibig, 2010), which may be induced by telling a lie.

In the guilty knowledge test, participants answered two types of questions: crime-relevant and crime-irrelevant questions. Crime-relevant questions asked whether they stole a wallet, whereas crime-irrelevant questions asked whether they stole other things such as jewelry. Guilty participants were to lie to crime-relevant questions. In this example, we analyzed participants' skin conductance level and pulse measured, while they were answering two questions: a crime-irrelevant question followed by a crime-relevant question. The skin conductance level and pulse were measured at 4.2 Hz for 10 seconds right after each question was posed. This yielded 84 time points in total: the first 42 time points for the crime-irrelevant question and the next 42 time points for the crime-relevant question.

We considered *skin conductance level* and *pulse* as two sets of predictor functions, and used the binary group (1 = guilty and 0 = innocent) as the scalar response variable. We hypothesized that the components of the two sets of predictor functions might predict the response variable. We also posited that the binary responses were generated from a binomial distribution under the number of trials = 1. We adopted the logit function as the canonical link.

We used B-spline functions to approximate the two sets of predictor functions and their weight functions, provided that the physiological measures involved non-periodic changes over time. The B-spline basis is a system of polynomials of prescribed order for each subinterval of S_k, which is split by so-called breakpoints or knots. To construct the B-spline basis, thus, it is necessary to specify the number of breakpoints and the order of polynomials. We set the number of breakpoints at 40 and the polynomial order at 4 (i.e., a cubic polynomial). Five-fold cross validation was applied for selecting the value of λ. Figure 11 displays the average minus log-

likelihood values against the common logarithms of different values of λ. The minimum of the average minus log-likelihood values was obtained at $\lambda = 10^6$. Thus, we decided $\lambda = 10^6$ as the smoothing parameter value for the example.

Insert Figure 11 about here

Given the value of λ, we applied the proposed method to fit the model to the data. Figure 12 displays the estimated weight functions for the two components. In the figure, the solid and dashed lines indicate the estimated weight functions and pointwise 95% confidence intervals, respectively. The horizontal dotted line indicates the zero line. A weight function is statistically significant at a particular point in time at the .05 level, if its 95% confidence intervals do not include a zero value in that point.

These weight functions showed the overall temporal characteristics of the two components. Panel (a) of Figure 12 exhibits the estimated weight function for *skin conductance level*. This weight function tended to increase monotonically over time. In particular, the skin conductance levels recorded over the first 42 time points were lower than those for the subsequent 42 time points. This suggests that when the crime-relevant question was asked, the participants generally showed higher skin conductance levels than when the crime-irrelevant question was posed. This weight function turned out to be statistically significant across most of the time points except for middle points.

As displayed in Panel (b) of Figure 12, the estimated weight function for *pulse* tended to increase until middle time points (i.e., between 40 and 50 time points) and to decrease for the remaining time points. This weight function was found statistically significant during the initial

period of asking the crime-relevant question. In particular, the weight value peaked at time point 43, which corresponded to the first record of pulse measured after the crime-relevant question was asked. This suggests that the level of pulse increased right after the participants were asked the crime-relevant question.

Insert Figure 12 about here

The loading estimates for the components of skin conductance level and pulse were 3.02 (se = 1.05, OR = 20.49) and 1.96 (se = .63, OR = 7.10), respectively. This indicates that both components had positive and statistically significant impacts on the binary response variable. Moreover, skin conductance level seems to have the greater influence on the response variable. The intercept estimate was 0.50 (se = 0.47).

6.2. The DVD Sales and Advertising Spending Data

The dataset used for the second example was prepared by linking DVD sales revenues collected from an on-line database (www.the-numbers.com) to advertising spending for DVD titles collected by a commercial advertising consulting company. This dataset included total weekly advertising spending in two different media (newspapers and network televisions) for 154 movie DVDs released in the US from 2006 to 2007, which were recorded continuously for 15 weeks (five weeks before DVD releases and ten weeks after releases). We used the two sets of weekly advertising spending records as predictor functions, and DVD sales calculated in the natural logarithmic scale as the scalar response variable. We hypothesized that the components of the two sets of predictor functions, called *newspaper advertising* and *network TV advertising*,

influenced the DVD sales. We examined the histogram and empirical quantile-quantile plot of the response variable to check whether it roughly followed a normal distribution. The response variable was not deviated substantially from a normal distribution. We thus assumed that observations on the DVD sales were generated from a normal distribution. We used the identity function as the canonical link.

We used B-splines to represent each set of predictor functions and its weight function, taking into account non-periodic changes of the spending measures over time. We chose the number of breakpoints and the polynomial order as 5 and 4, respectively, for the construction of the B-spline functions. We applied five-fold cross validation for selecting the value of λ. Figure 13 displays the average minus log-likelihood values against the common logarithms of different values of λ. As shown in the figure, the minimum value was achieved at $\lambda = 10^2$. Thus, we chose $\lambda = 10^2$ for this example.

Insert Figure 13 about here

We then applied the proposed method to fit the model to the data. Figure 14 displays the estimated weight functions for the two components. In the figure, again, the solid and dashed lines indicate the estimated weight functions and pointwise 95% confidence intervals, respectively. The horizontal dotted line indicates the zero line.

As shown in Panel (a) of Figure 14, the weight function for *newspaper advertising* appeared to increase gradually until the 10^{th} week and to decrease for the remaining weeks. However, this weight function turned out to be statistically significant only between the eighth and eleventh weeks, suggesting a non-negligible level of spending on newspaper advertising

during the particular period. As shown in Panel (b) of Figure 14, the weight function values for *network TV advertising* appeared to peak at week 5 and to decrease for the subsequent weeks. Moreover, the weight function for network TV advertising was statistically significant from weeks 3 to 8. This indicates that network TV advertising was concentrated on three weeks prior to and three weeks after the release week.

Insert Figure 14 about here

The loading estimates for the components of newspaper advertising and network TV advertising were .13 (se = .05) and .52 (se = .05), respectively. This indicates that newspaper advertising and network TV advertising had positive and statistically significant impacts on the DVD sales. The intercept estimate was 17.30 (se = .05).

7. Concluding Remarks

We proposed an extension of functional extended redundancy analysis to the framework of generalized linear models. This extension permits studying the effects of components of multiple sets of predictor functions on scalar responses arising from diverse types of distributions in the exponential family. A penalized log-likelihood criterion was developed that entailed a penalty term for controlling for roughness of weight functions. By adopting a basis expansion approach to approximating predictor and weight functions, maximization of this criterion becomes similar to conventional maximum-likelihood estimation procedures for univariate response data.

We conducted simulation studies to investigate the parameter recovery capability of the proposed method. In general, the method resulted in parameter estimates that were close to their parameters for normal and binomial data. Moreover, it produced more accurate parameter estimates than a two-step approach that applied a functional principal components analysis and a generalized linear model sequentially. We also applied the proposed method to two real datasets, each of which had a different type of response data. For both applications, the method was of use in summarizing overall, temporal characteristics of each set of predictor functions as well as its influence on a response variable.

We may further extend the proposed method to improve its data-analytic capability and the scope of its applicability. For example, the proposed method estimates parameters by maximizing a penalized log-likelihood criterion. This maximum-likelihood estimation requires that the parametric form of the distribution by which observations are generated be completely known so as to construct the likelihood. However, at times, it may be difficult to have prior theory or sufficient knowledge about the entire parametric form of the distribution for response data. In such situations, we may relax the full specification of the probability distributions underlying response data by replacing likelihoods with quasi-likelihoods (Heyde, 1997; McCullagh, 1983; Wedderburn, 1974). Quasi-likelihoods behave similarly to likelihoods, but require information only on the mean and variance of observations.

In addition, we may combine some clustering technique with the proposed method in a unified manner in order to capture group-wise heterogeneity of observations. The proposed method estimates parameters by aggregating the data across observations under the implicit assumption that all observations arise from a single homogenous group. This aggregate-sample analysis may not be suitable when there exist subgroups of observations that elicit distinct

characteristics. To account for such group-wise heterogeneity, we may develop a finite mixture approach (e.g., McLachlan & Peel, 2000) to the proposed method. In particular, this finite mixture extension will combine the proposed method with finite mixture generalized linear models (Wedel & DeSarbo, 1995). The finite mixture extension involves likelihood-based estimation procedures, and again requires the complete specification of the probabilistic distribution by which the data at hand are generated. If it is untenable to fully understand the entire mechanism of data generation, we may consider combining the proposed method with fuzzy clusterwise quasi-likelihood estimation procedures (Hwang & Tomiuk, 2010).

As stated earlier, it is consistent with standard GLM that the proposed method focuses on scalar responses arising from a univariate distribution in the exponential family. Nonetheless, it may be worthwhile to extend the method to analyze multivariate response data. Yee and Wild (1999) proposed vector generalized linear models which deal with a broad range of multivariate, exponential-family response data. We may incorporate their approach into the proposed method.

References

Dauxois, J., Pousse, A., & Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, *12*, 136-154.

Davidian, M., Lin, X., & Wang, J.-L. (2004). Introduction: emerging issues in longitudinal and functional data analysis. *Statistica Sinica*, *14*, 613-614.

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51, 1-28.

Ferraty, F. (2011). *Recent advances in functional data analysis and related topics*. New York: Springer.

Ferraty, F., Laksaci, A., Tadj, A., & Philippe Vieu, P. (2011). Kernel regression with functional response. *Electronic Journal of Statistics*, 5, 159-171.

Ferraty, F., Mas, A., & Vieu, P. (2007). Nonparametric regression on functional data: inference and practical. *Australian & New Zealand Journal of Statistics*, 49, 267-286.

Ferraty, F., & Romain, Y. (2011). *The Oxford handbook of functional data analysis*. Oxford: University Press.

Ferraty, F., Van Keilegom, I., & Vieu, P. (2012). Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, 109, 10-28.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice.* New York: Springer.

González-Manteiga, W., & Vieu, P. (2007). Statistics for functional data. *Computational Statistics & Data Analysis*, *51*, 4788-4792.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternative (with discussion). *Journal of the Royal Statistical Society B*, *46*, 149-192.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York: Springer.

Heyde, C. C. (1997). *Quasi-likelihood and its Applications. A general approach to optimal parameter estimation*. New York: Springer-Verlag.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Application to nonorthogonal problems. *Technometrics*, *12*, 69-82.

Hwang, H. (2009). Regularized generalized structured component analysis. *Psychometrika*, *74*, 517-530.

Hwang, H., Suk, H. W., Lee, J.-H., Moskowitz, D. S., & Lim, J. (2012). Functional extended redundancy analysis. *Psychometrika*, *77*, 524-542.

Hwang, H., & Tomiuk, M. A. (2010). Fuzzy clusterwise quasi-likelihood generalized linear models. *Advances in Data Analysis and Classification*, *4*, 255-270.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*, 394-421.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics*, *41*, 191-201.

Lee, A., & Silvapulle, M. (1988). Ridge estimation in logistic regression. *Communications in Statistics, Simulation and Computation*, *17*, 1231-1257.

Lian, H. (2011). Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics*, 5, 31-40.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, *11*, 59-67.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall/CRC.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A*, *135*, 370-384.

Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society B*, *53*, 539-572.

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and Matlab*. New York: Springer.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis* (1st ed.). New York: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case Studies*. New York: Springer.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Rice, J. A. (2004). Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, *14*, 631-647.

Rice, J. A., & Silverman, B. W. (1991). Estimating the mean and covariance structure non-parametrically when the data are curves. *The Journal of the Royal Statistical Society B*, *53*, 233-243.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *Journal of the Royal Statistical Society B*, *23*, 469-475.

Takane, Y., & Hwang, H. (2005). An extended redundancy analysis and its applications to two practical examples. *Computational Statistics and Data Analysis*, *49*, 785-808.

Valderrama, M. J. (2007). An overview to modelling functional data. *Computational Statistics*, *22*, 331-334.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439-447.

Wedel, M., & DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, *12*, 21-55.

Yee, T. W. & Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, *3*, 15-41.

Yee, T. W., & Wild, C. J. (1999). Vector generalized additive models. *Journal of the Royal Statistical Society B*, *58*, 481-493.

Table 1. The mean square errors (MSE), squared biases (SBIAS), and variances (VAR) of loading estimates obtained from the proposed method (GFERA) and two-step approach (Two-step) at different sample sizes for normal data.

		a_0			a_1			a_2		
	N	MSE	SBIAS	VAR	MSE	SBIAS	VAR	MSE	SBIAS	VAR
GFERA	50	0.0127	0.0003	0.0124	0.0133	0.0000	0.0132	0.0169	0.0009	0.0160
	100	0.0059	0.0003	0.0056	0.0069	0.0000	0.0068	0.0063	0.0001	0.0062
	200	0.0029	0.0000	0.0029	0.0024	0.0000	0.0023	0.0038	0.0000	0.0038
	500	0.0011	0.0000	0.0011	0.0012	0.0000	0.0011	0.0011	0.0000	0.0011
Two-step	50	0.0127	0.0003	0.0124	0.0625	0.0508	0.0118	0.0182	0.0055	0.0128
	100	0.0059	0.0003	0.0056	0.1697	0.1642	0.0055	0.0417	0.0355	0.0062
	200	0.0029	0.0000	0.0029	0.0406	0.0382	0.0025	0.2403	0.2375	0.0028
	500	0.0011	0.0000	0.0011	0.0617	0.0606	0.0012	0.3666	0.3656	0.0010

Table 2. The mean square errors (MSE), squared biases (SBIAS), and variances (VAR) of loading estimates obtained from the proposed method (GFERA) and two-step approach (Two-step) at different sample sizes for binomial data.

		a_0			a_1			a_2		
	N	MSE	SBIAS	VAR	MSE	SBIAS	VAR	MSE	SBIAS	VAR
GFERA	50	0.1437	0.0072	0.1366	0.2432	0.0355	0.2077	0.1655	0.0069	0.1587
	100	0.0543	0.0015	0.0529	0.1140	0.0086	0.1055	0.1330	0.0220	0.1111
	200	0.0279	0.0030	0.0249	0.0482	0.0004	0.0478	0.0435	0.0015	0.0420
	500	0.0117	0.0002	0.0115	0.0125	0.0007	0.0118	0.0159	0.0013	0.0146
Two-step	50	0.0946	0.0027	0.0919	0.1771	0.0860	0.0910	0.3821	0.3489	0.0332
	100	0.0325	0.0001	0.0324	0.1909	0.1622	0.0288	0.2277	0.1795	0.0482
	200	0.0182	0.0000	0.0182	0.0670	0.0453	0.0216	0.2929	0.2715	0.0214
	500	0.0092	0.0014	0.0078	0.1421	0.1350	0.0071	0.5109	0.5074	0.0034

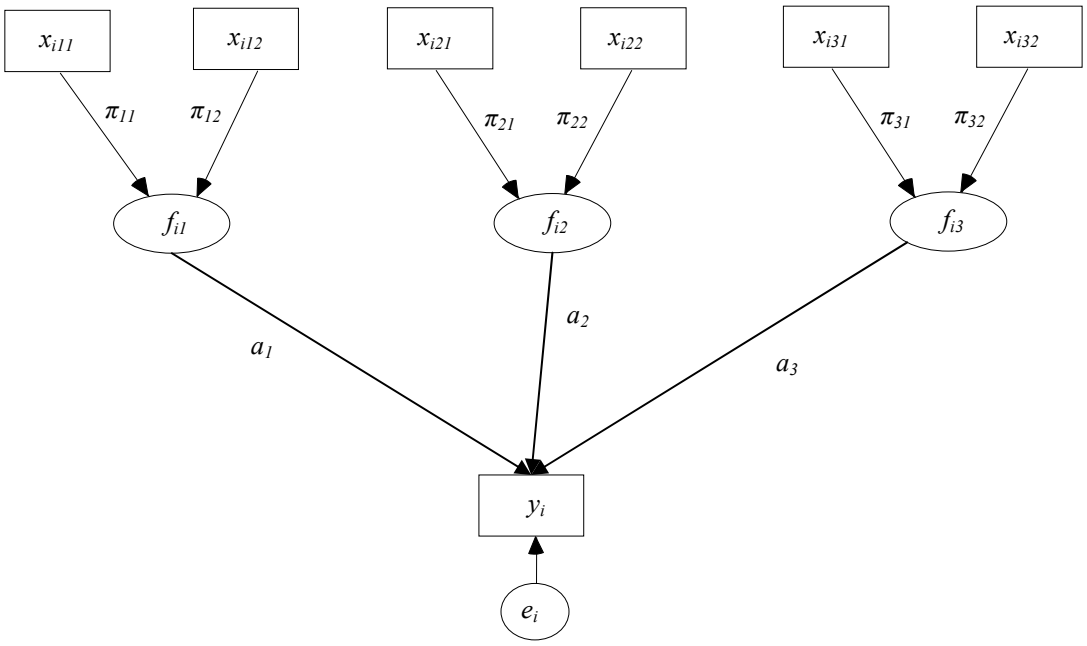
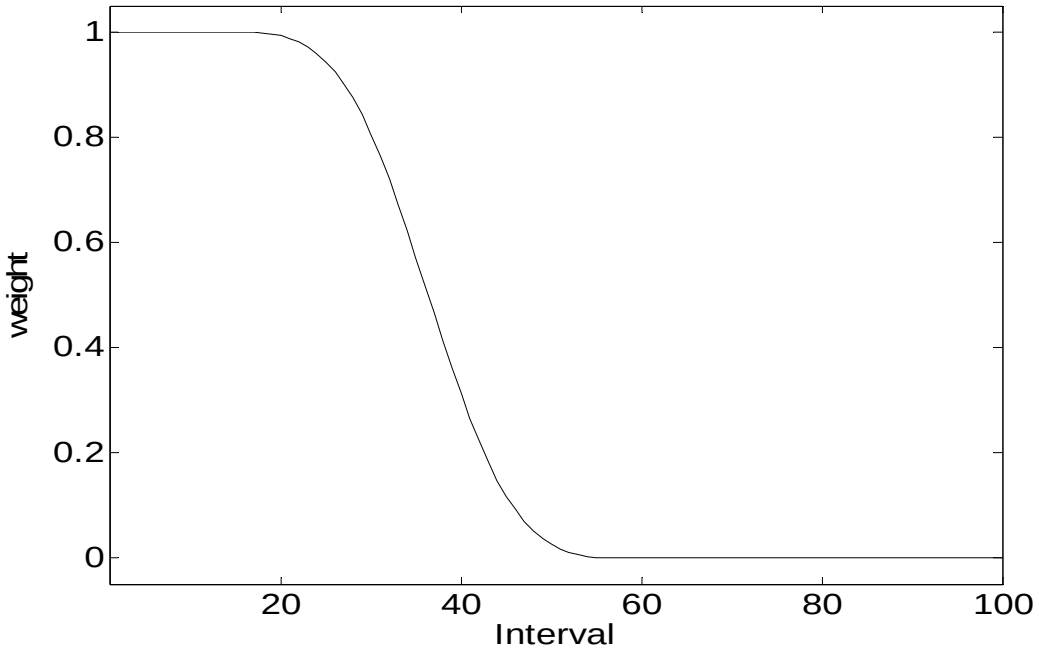


Figure 1. A path diagram of an extended redundancy analysis model. In the diagram, y_i is the response variable value measured on the ith observation ($i = 1,..., N$); x_{ikp} is the pth variable value in the kth set of predictor variables on the ith observation (k = 1, 2, 3; p = 1, 2); π_{kp} is a component weight assigned to x_{ikp}; f_{ik} is the ith component score of the kth set of predictor variables; a_k a component loading relating f_{ik} to y_i; and e_i is the residual value of y_i.

(a)



(b)

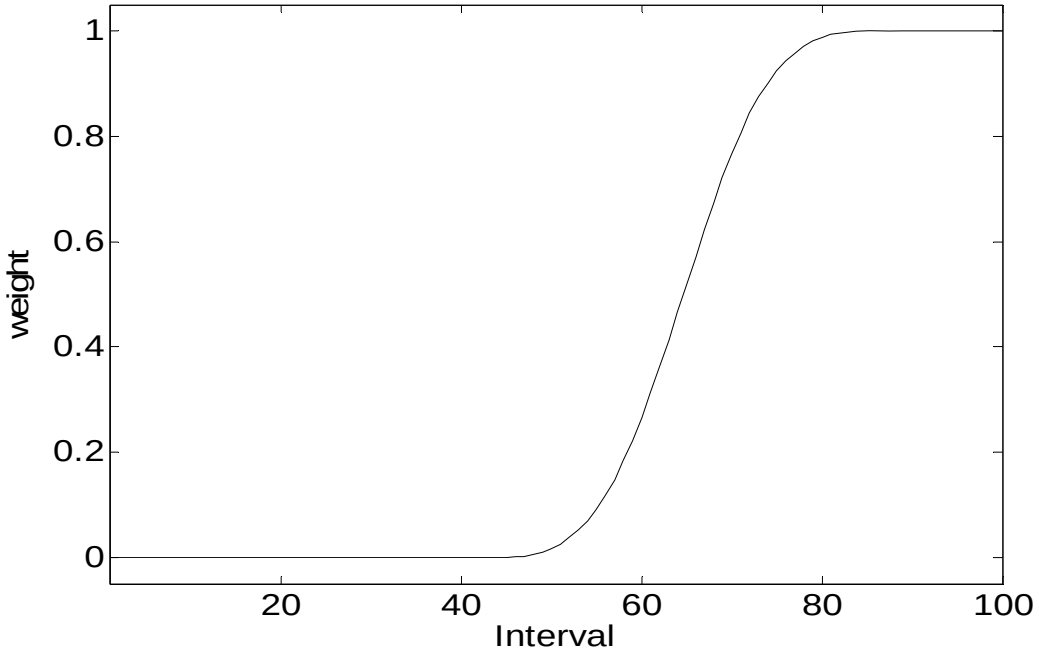
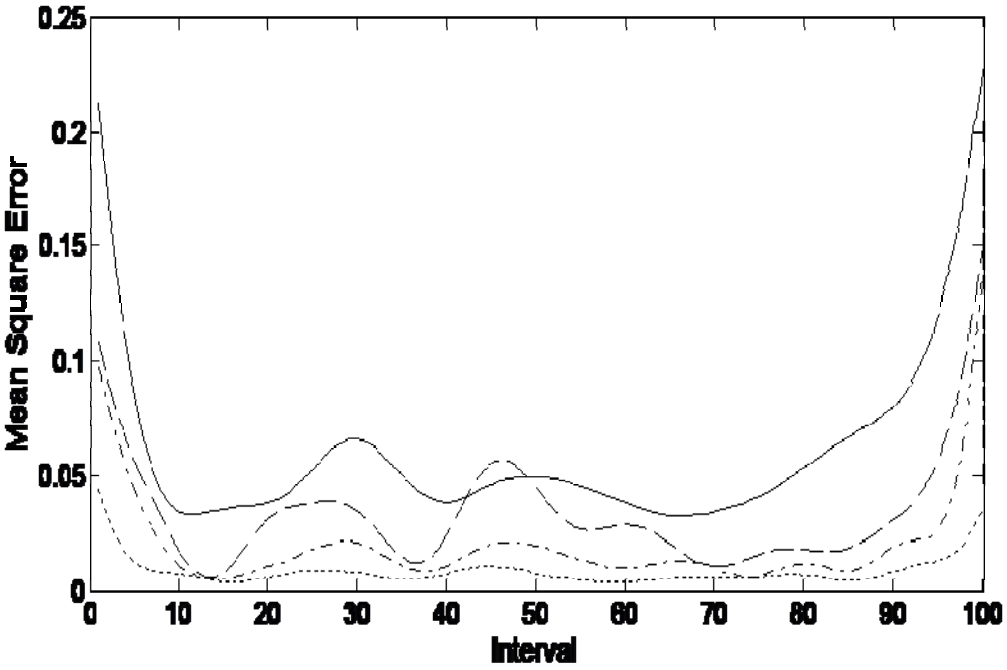


Figure 2. The two weight functions used in the simulation studies: (a) weight function 1 and (b) weight function 2.

(a)



(b)

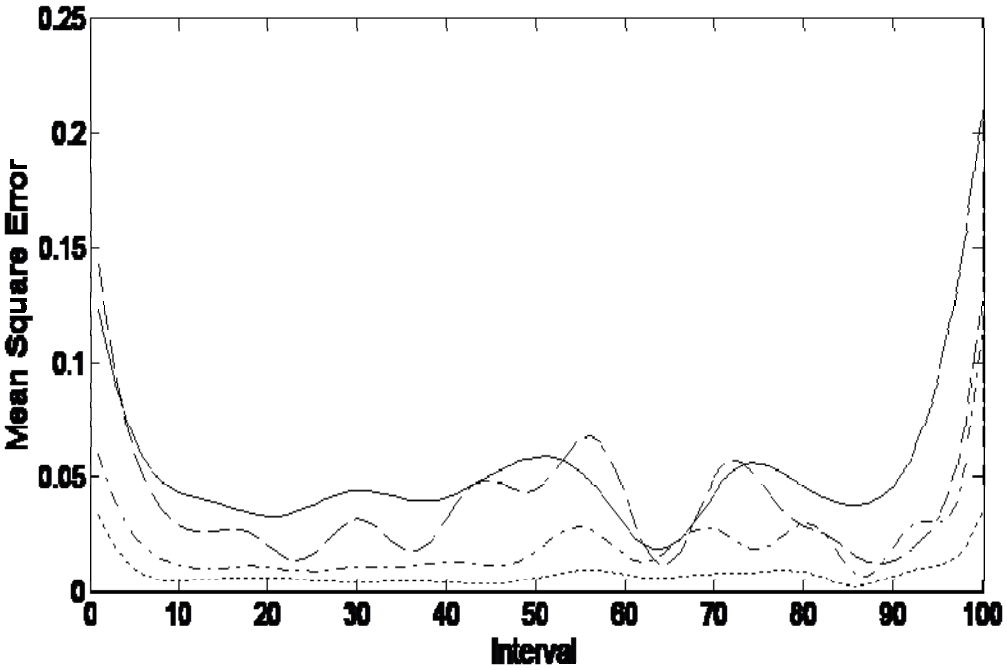
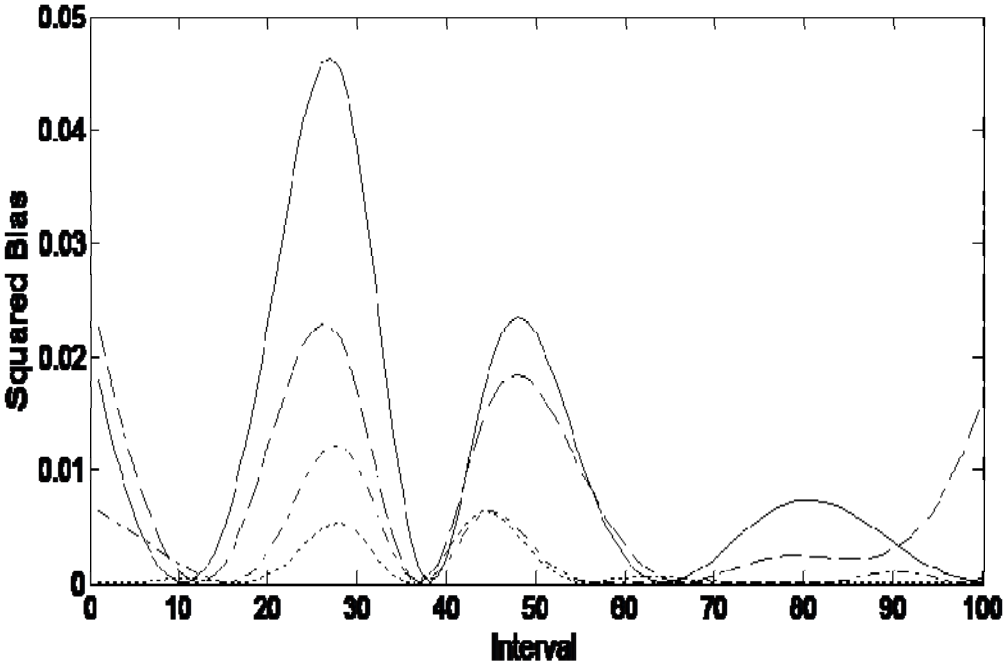


Figure 3. The mean square errors of estimated two weight functions obtained from the proposed method for normal data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

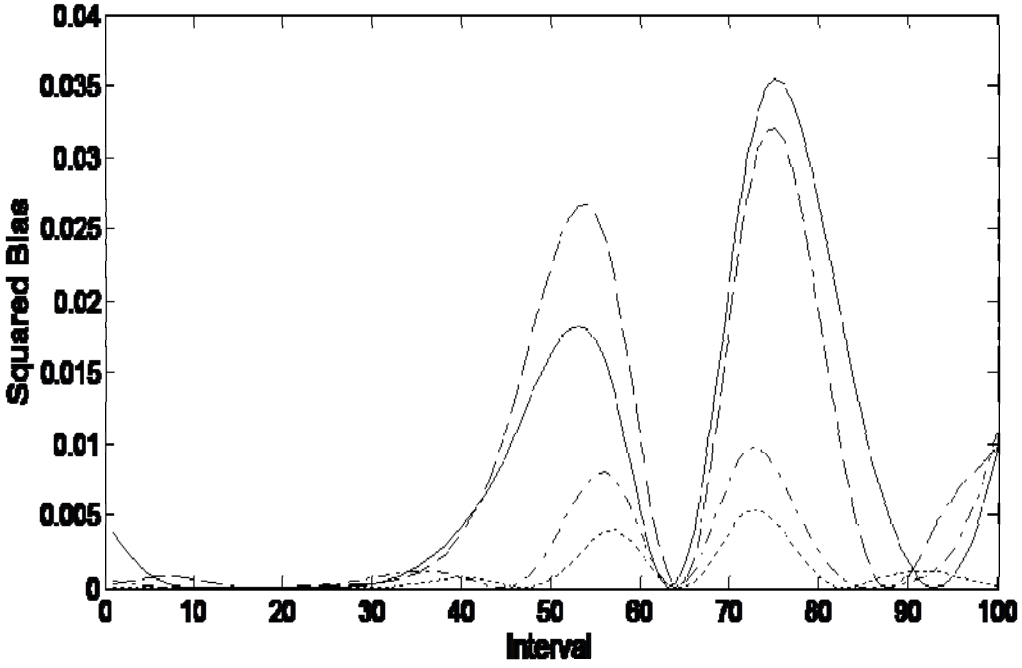
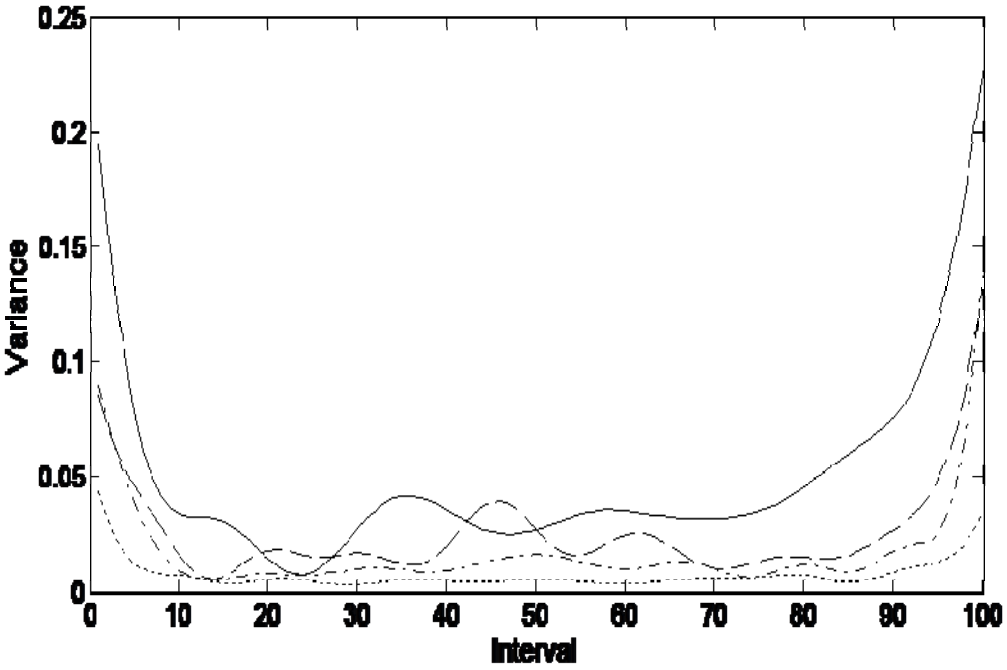


Figure 4. The squared biases of estimated two weight functions obtained from the proposed method for normal data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

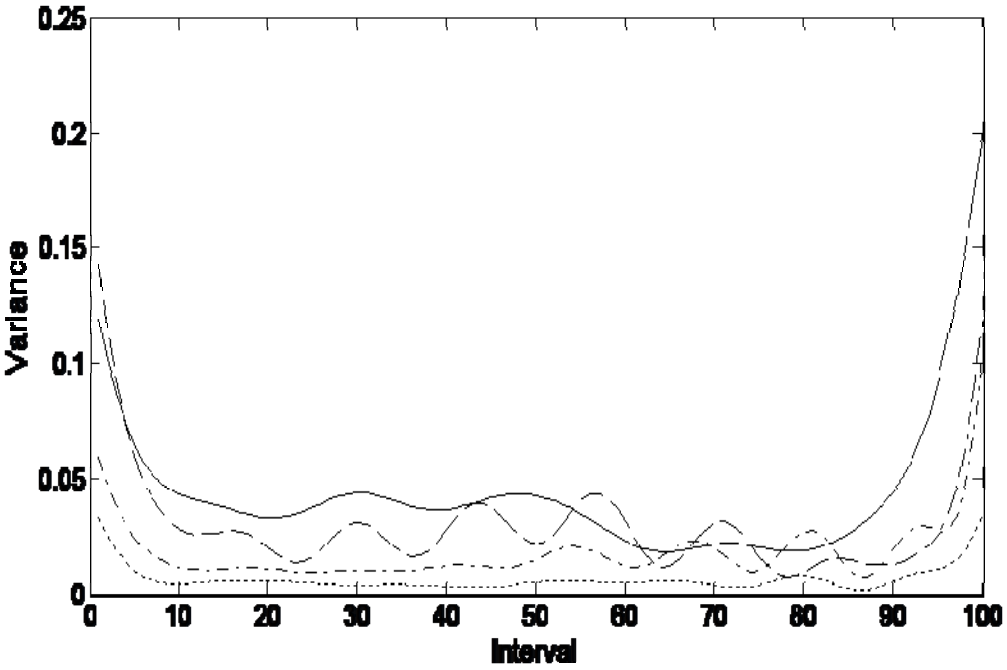
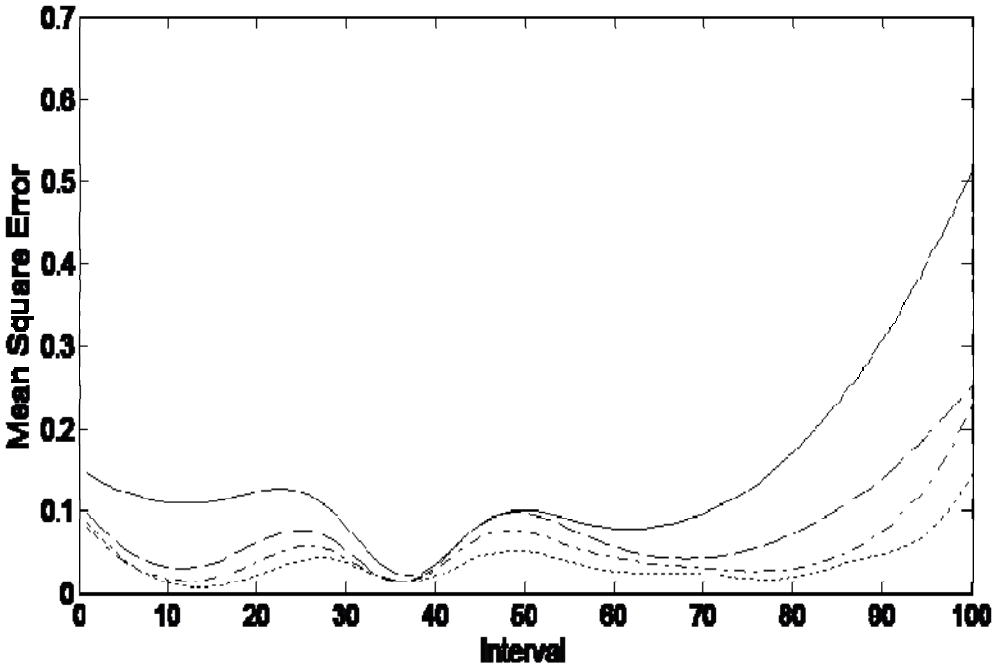


Figure 5. The variances of estimated two weight functions obtained from the two-step approach for normal data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

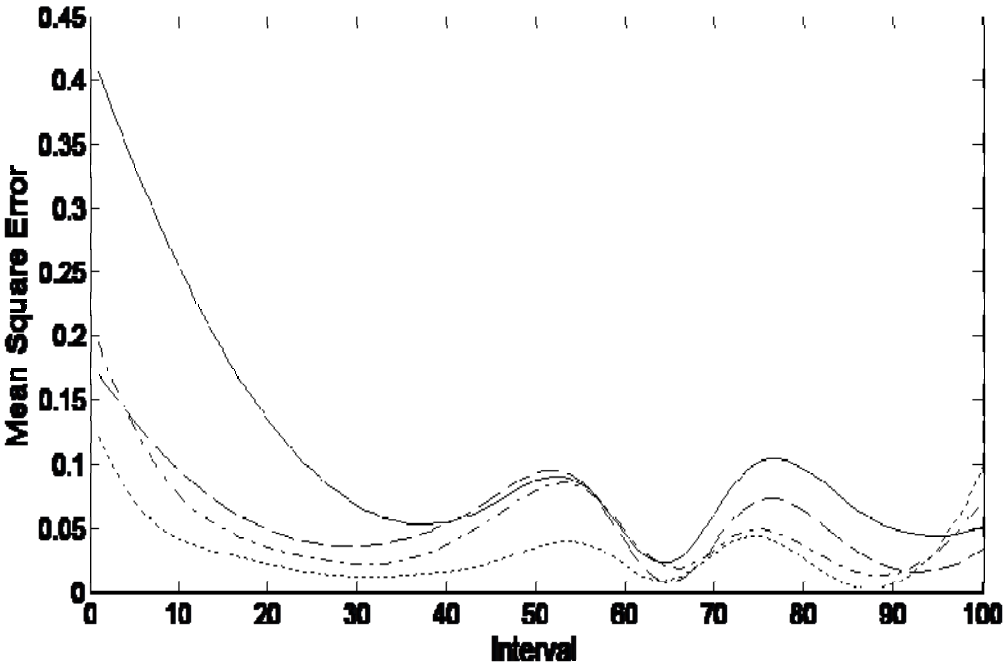
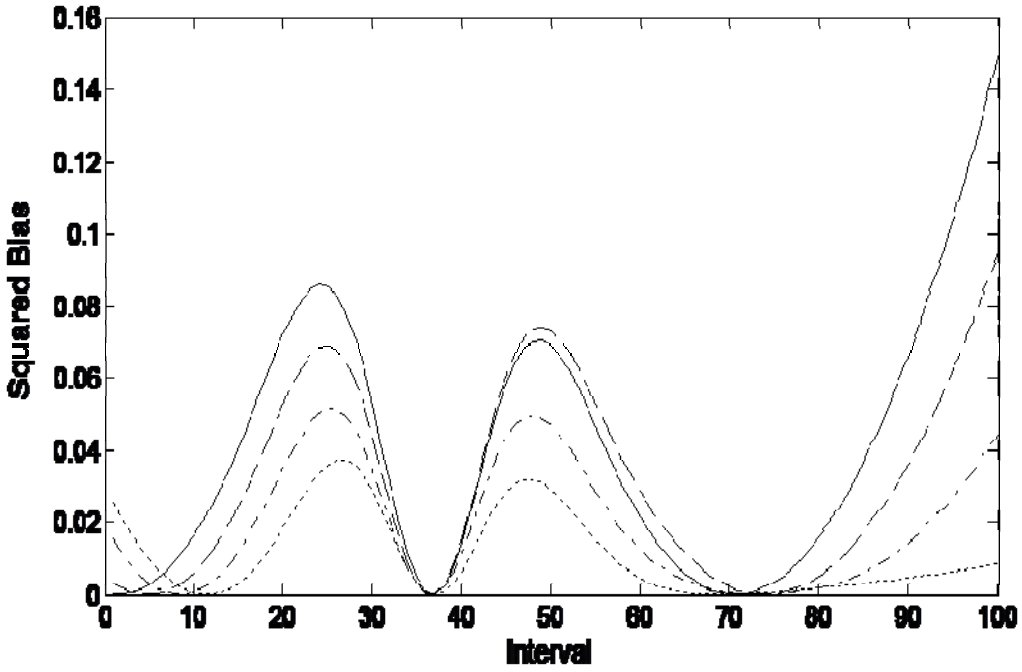


Figure 6. The mean square errors of estimated two weight functions obtained from the proposed method for binomial data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

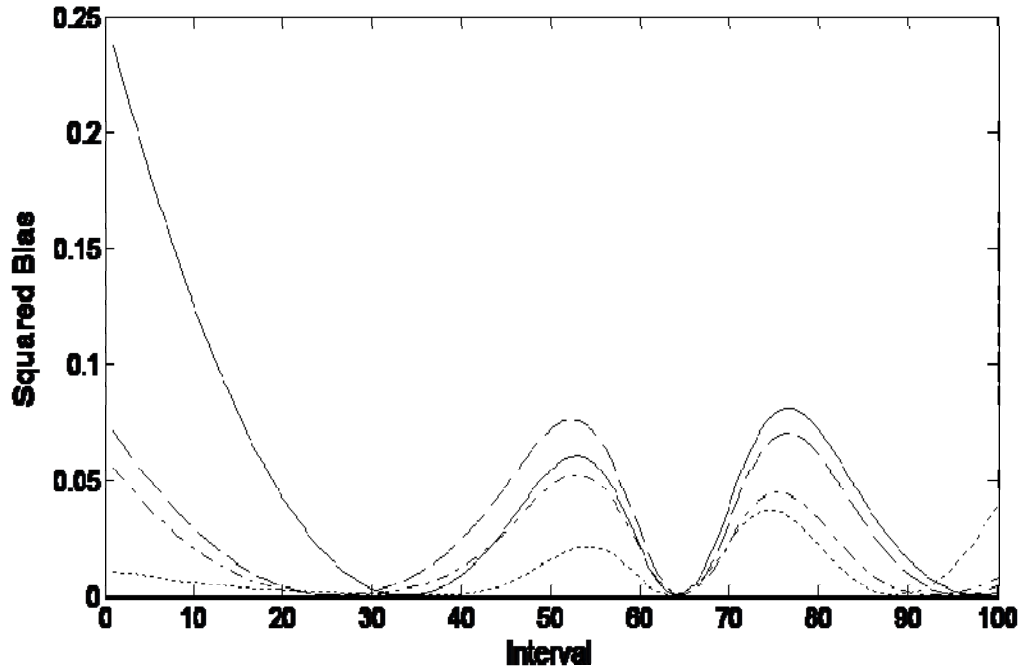
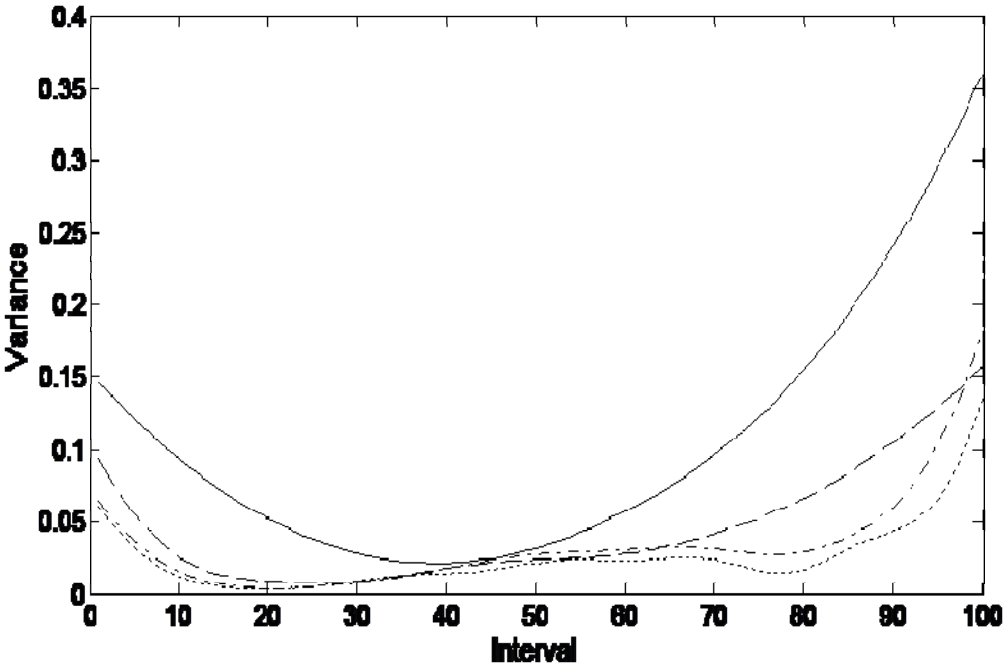


Figure 7. The squared biases of estimated two weight functions obtained from the proposed method for binomial data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

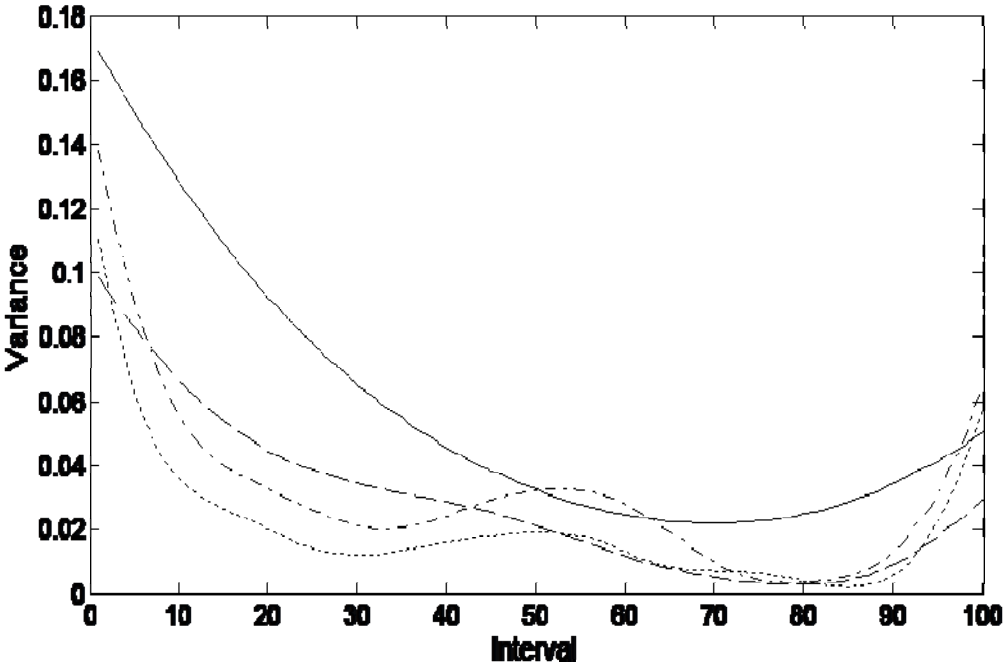
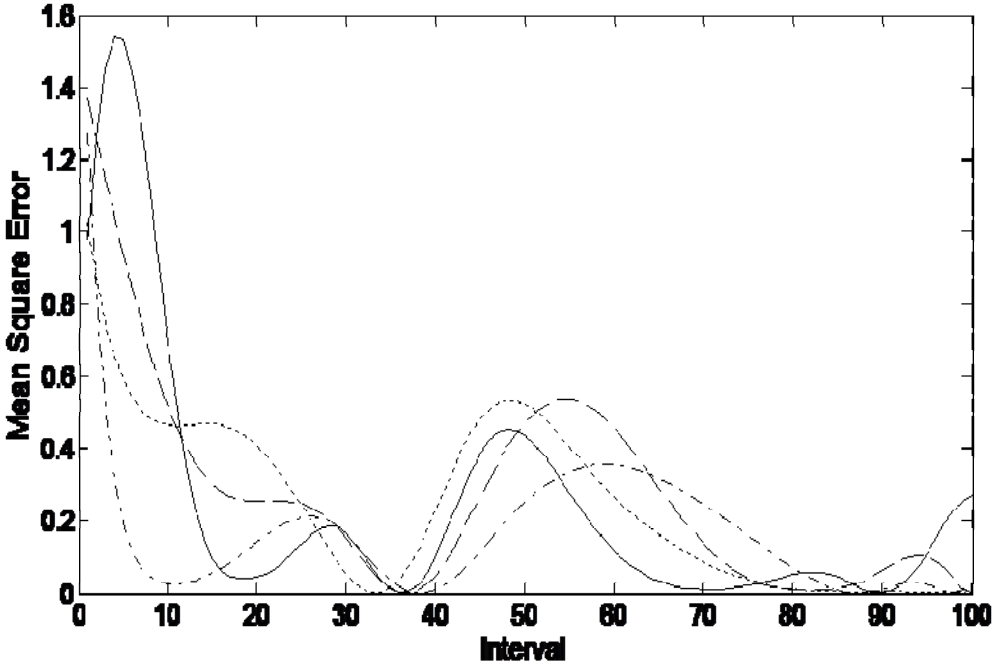


Figure 8. The variances of estimated two weight functions obtained from the proposed method for binomial data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

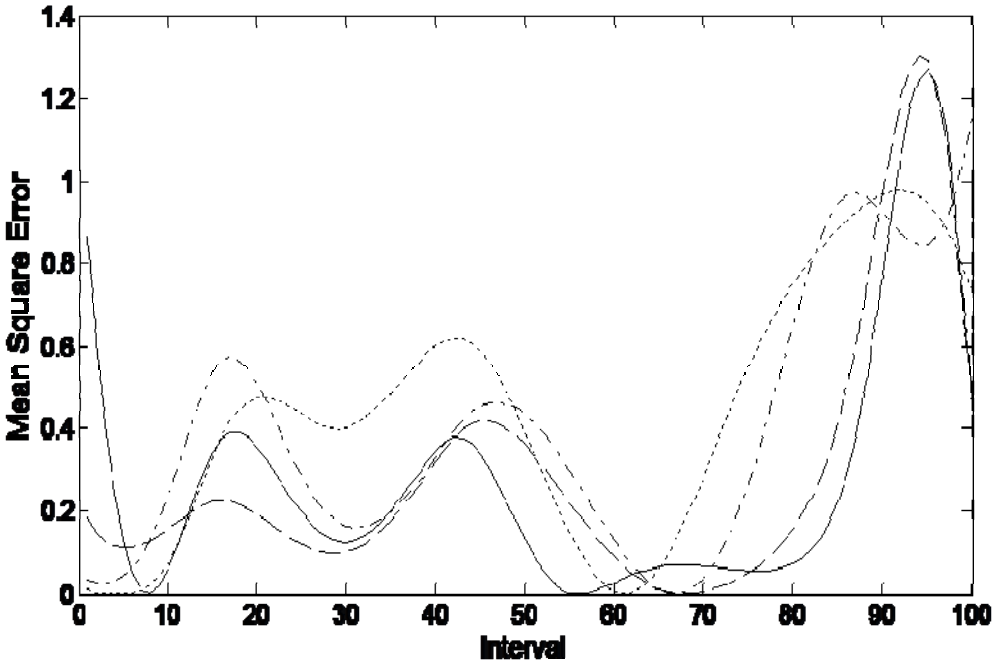
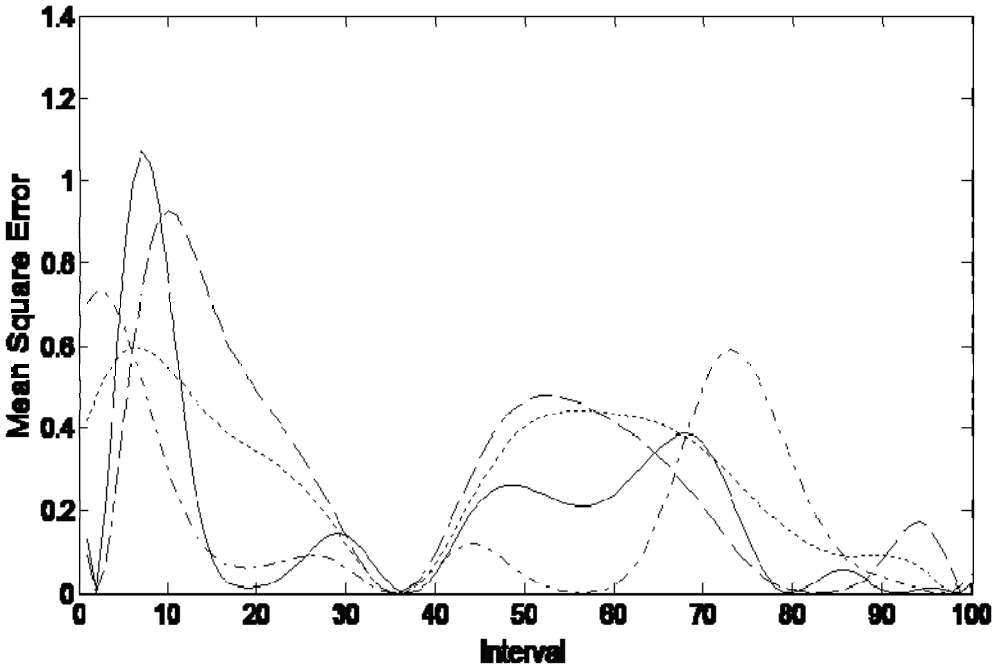


Figure 9. The mean square errors of estimated two weight functions obtained from the two-step approach for normal data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

(a)



(b)

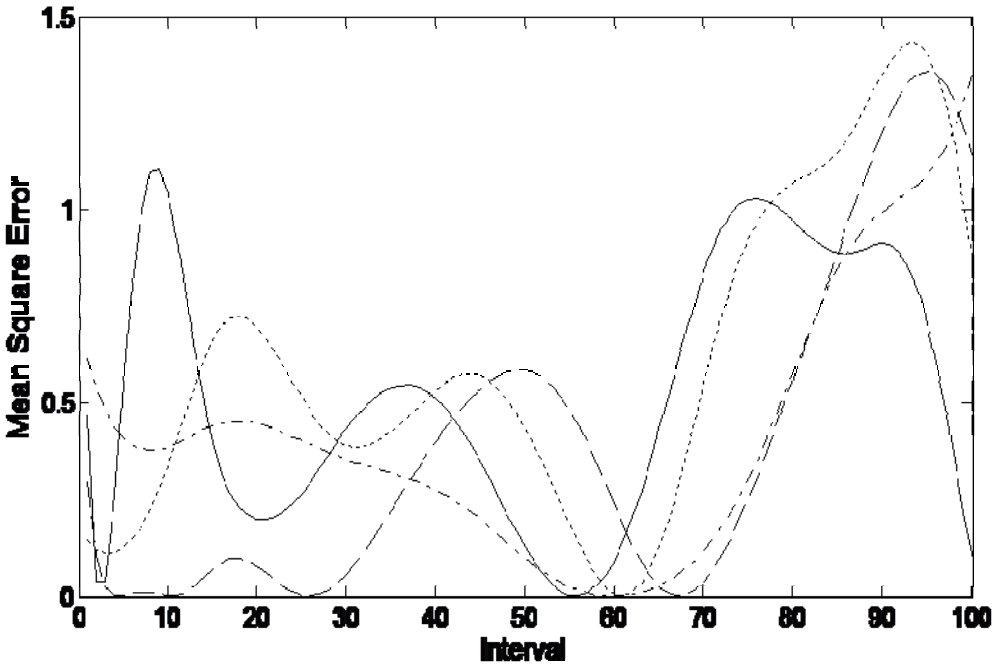


Figure 10. The mean square errors of estimated two weight functions obtained from the two-step approach for binomial data at $N = 50$ (solid line), 100 (dashed line), 200 (dash-dotted line), and 500 (dotted line).

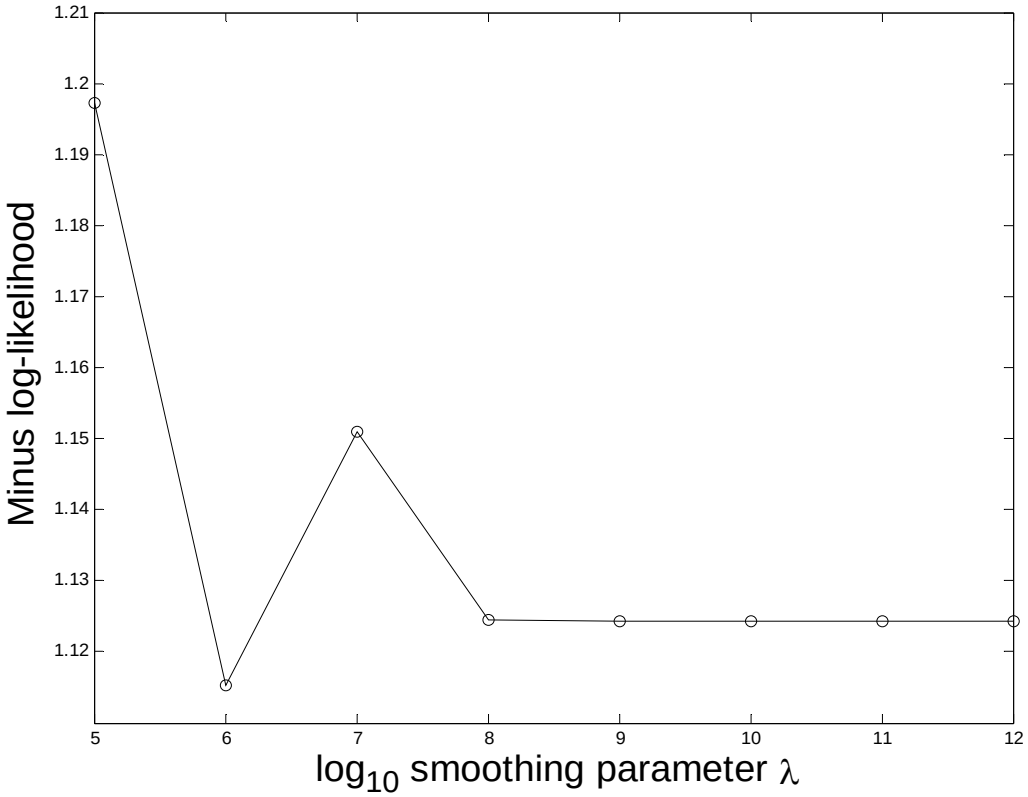
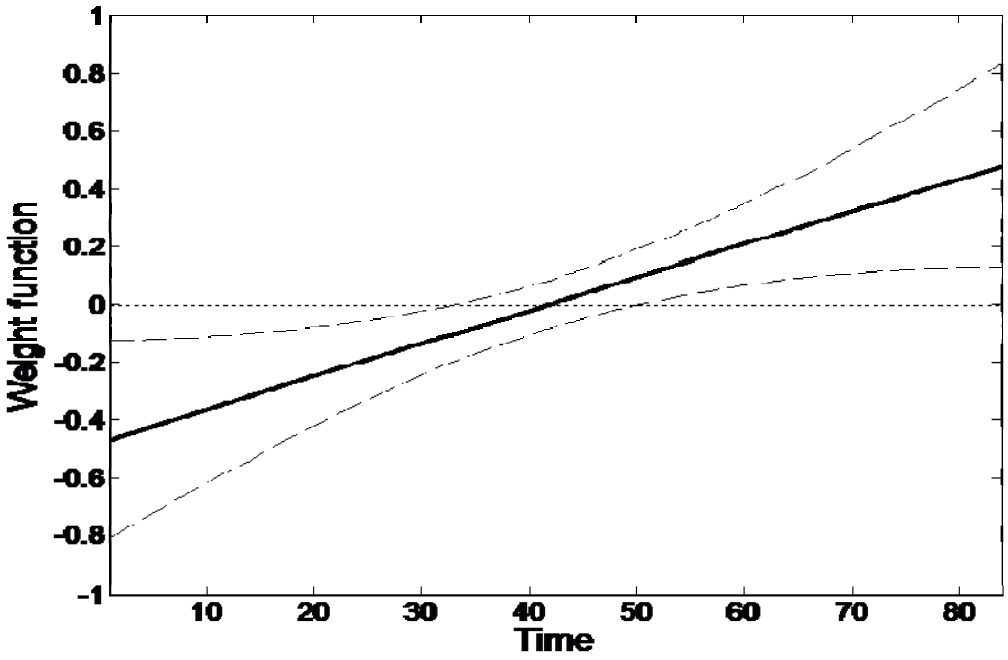


Figure 11. The average minus log-likelihood values against the common logarithms of different values of λ for the lie detection data.

(a)



(b)

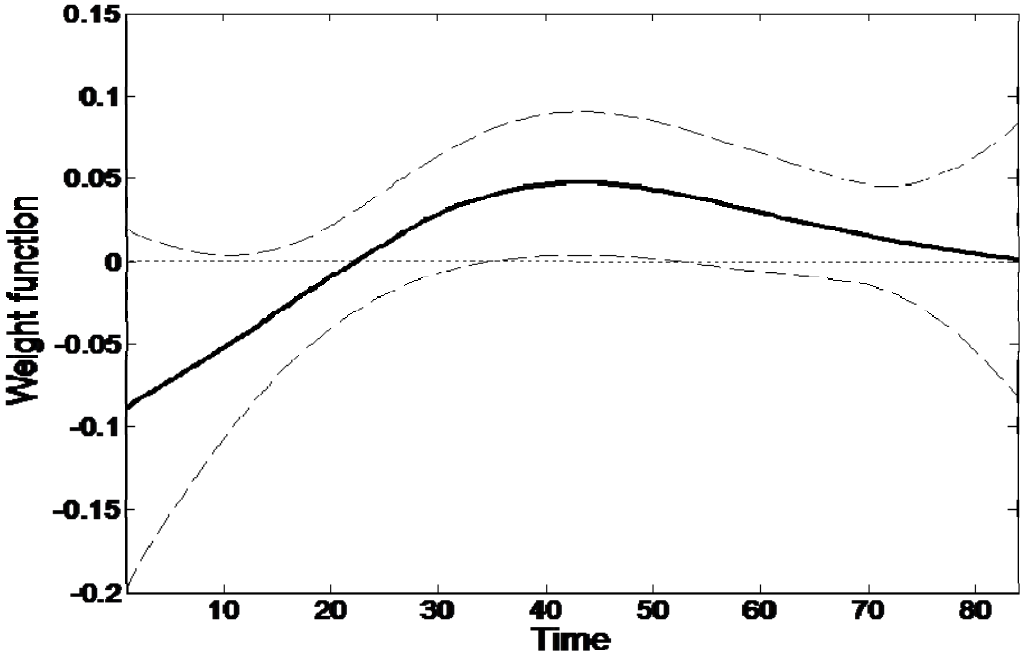


Figure 12. The estimated weight functions (thick solid lines) and their pointwise 95% confidence intervals (dashed lines) for (a) skin conductance level and (b) pulse. The horizontal dotted line indicates the zero line.

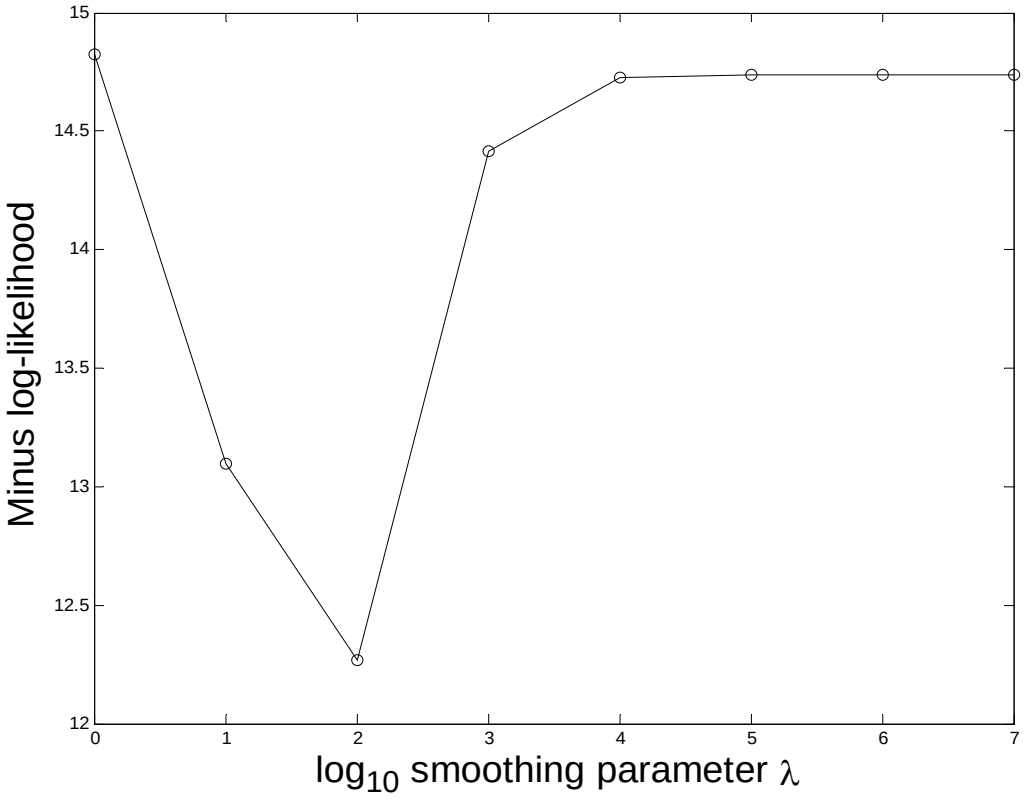
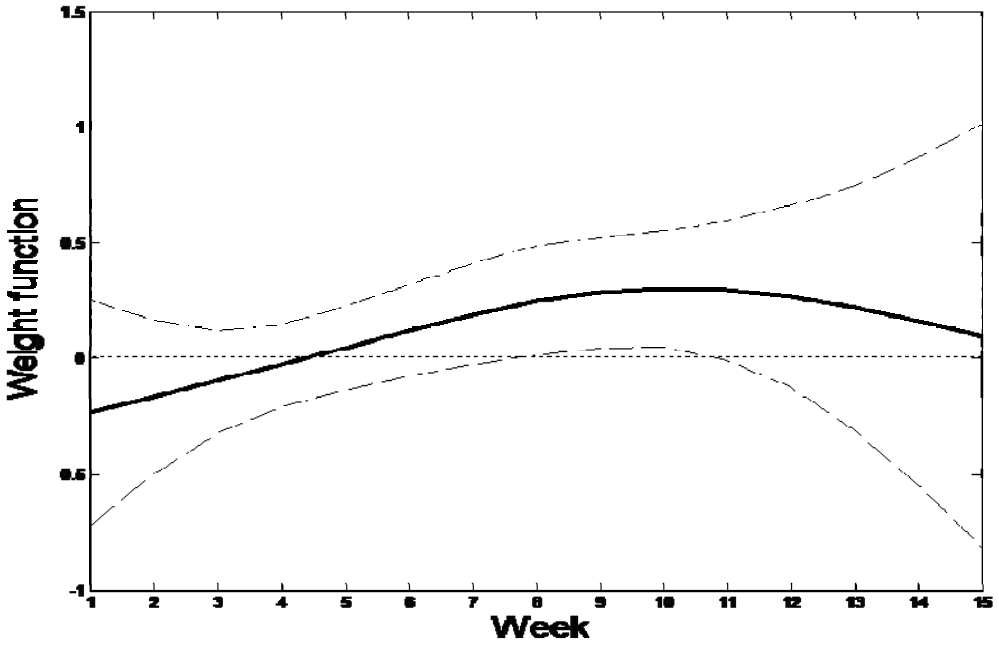


Figure 13. The average minus log-likelihood values against the common logarithms of different values of λ for the DVD sales and advertising spending data.

(a)



(b)

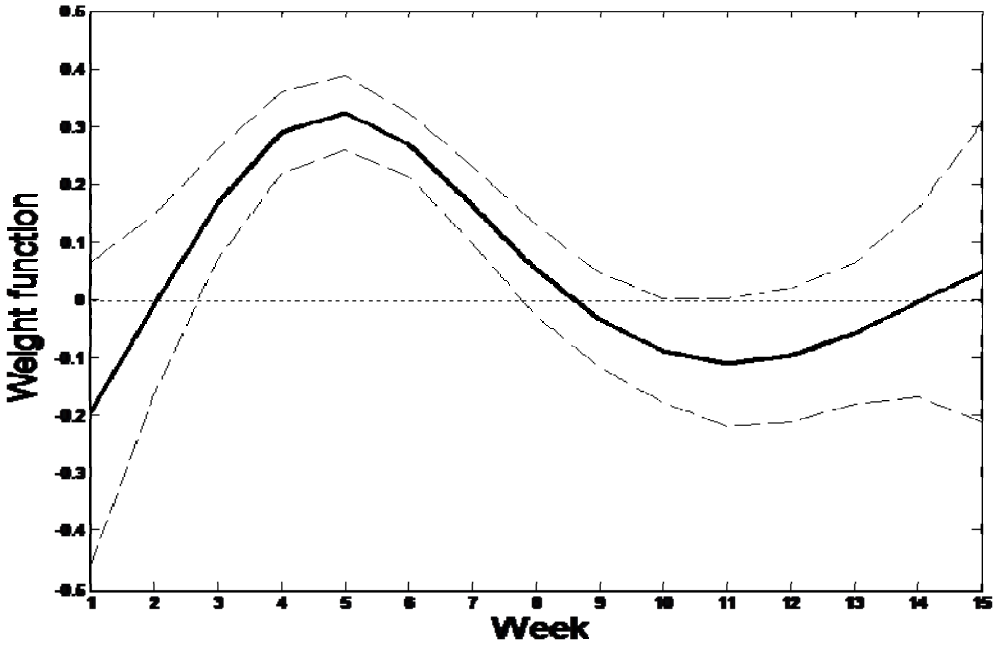


Figure 14. The estimated loading functions (thick solid lines) and their pointwise 95% confidence intervals (dashed lines) for (a) newspaper advertising and (b) network TV advertising. The horizontal dotted line indicates the zero line.