Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 2285–2303
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

On Stein's lemma, dependent covariates and functional monotonicity in multi-dimensional modeling

Chunming Zhang[a,*], Jialiang Li[b], Jingci Meng[a]

[a] *Department of Statistics, University of Wisconsin, Medical Sciences Center, 1300 University Avenue, Madison, WI 53706, United States*
[b] *Department of Statistics & Applied Probability, National University of Singapore, Singapore 117546, Singapore*

Received 14 April 2006
Available online 14 February 2008

Abstract

Tracking the correct directions of monotonicity in multi-dimensional modeling plays an important role in interpreting functional associations. In the presence of multiple predictors, we provide empirical evidence that the observed monotone directions via parametric, nonparametric or semiparametric fit of commonly used multi-dimensional models may entirely violate the actual directions of monotonicity. This breakdown is caused primarily by the dependence structure of covariates, with negligible influence from the bias of function estimation. To examine the linkage between the dependent covariates and monotone directions, we first generalize Stein's Lemma for random variables which are mutually independent Gaussian to two important cases: dependent Gaussian, and independent non-Gaussian. We show that in both two cases, there is an explicit one-to-one correspondence between the monotone directions of a multi-dimensional function and the signs of a deterministic surrogate vector. Moreover, we demonstrate that the second case can be extended to accommodate a class of dependent covariates. This generalization further enables us to develop a de-correlation transform for arbitrarily dependent covariates. The transformed covariates preserve modeling interpretability with little loss in modeling efficiency. The simplicity and effectiveness of the proposed method are illustrated via simulation studies and real data application.

AMS 2000 subject classifications: primary 62F30; 62H20; secondary 62E15; 62H10

Keywords: Additive model; Nonparametric regression; Partially monotone function; Similarly ordered; Stein's Lemma; Support vector machine

* Corresponding author.
E-mail addresses: cmzhang@stat.wisc.edu (C. Zhang), stalj@nus.edu.sg (J. Li), mengj@stat.wisc.edu (J. Meng).

0047-259X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.02.018

1. Introduction

Monotone functions are central to order theory. In statistical applications, some notable shape property, such as monotonicity, is particularly useful for interpreting functional associations between a response variable and predictor variables. For example, in the environmental study of ozone data [12], one may naturally ask whether the measurement of ozone concentration is monotone increasing in temperature and monotone decreasing in wind speed. In socio-economical studies, much theoretical and empirical literature predicts that wages increase with age and education [17]. In many other similar applications arising from biomedical and engineering studies, one would postulate a statistical model such that the modeling function of covariates preserves the isotonic assumptions on a subset of the covariates. Most importantly, one would expect the monotonicity property to be inherited by data-based estimates of the function.

Nevertheless, in the presence of multiple covariates, preserving the functional monotonicity, from finite-sample estimates, will be much more challenging than in the case of a univariate predictor. As will be seen, the observed directions of monotonicity from parametric or nonparametric estimates may deviate significantly from the actual directions of monotonicity. In statistical literature, while many useful asymptotic results have been established for the consistency of parametric and nonparametric estimates, the results could not directly explain the discrepancy between the observed directions of monotonicity and the actual directions of monotonicity.

Fig. 1 illustrates the extent to which the observed functional monotonicity in two covariates departs from the actual monotonicity. There, the response variable follows a bivariate additive model in which the true regression function is monotone increasing in X_1 and monotone decreasing in X_2. For random samples generated from this model, the component functions are estimated by the local linear backfitting procedure. The asymptotic normality of the resulting estimates has been shown in [22]. To evaluate whether the fitted curves correctly reveal the actual isotonic directions, three types of dependence structure between covariates X_1 and X_2 are examined. In the first case where X_1 and X_2 are dependent Gaussian as well as in the second case where they are independent non-Gaussian, both the observed directions of monotonicity from the function estimates agree with the actual monotone directions. Curiously, in the third case where X_1 and X_2 follow a mixture of bivariate normal distributions, it is apparent that the estimated curve with respect to X_2 entirely violates the assumption of monotone decreasing. Fig. 1 makes it evident that the structure and magnitude of dependence between covariates need to be taken into account when a multi-dimensional model is employed.

The empirical evidence provided from the above third case is not pathological. Indeed, similar phenomena arise from many popular multi-dimensional models, as will be exemplified by extensive studies in Section 4. This is an interesting problem of both theoretical and practical importance. However, addressing the issue raised above is a nontrivial task. In a parametric multiple linear regression model, this phenomenon may be qualitatively attributed to collinearity. Nonetheless, for non-parametric and semi-parametric modeling, little published information exists to explain this empirical result. Hence a more careful, unified and quantitative study is needed. We will investigate the issues of how the dependence affects monotonicity and how the lack of monotonicity could be reduced in statistical analysis with multivariate covariates.

There is diverse and extensive literature addressing monotonicity. These include [8,5,17,24], among many others. Most of the published results are confined to one or two dimensions and, in those circumstances, the majority of the function estimates are monotone and only a small fraction violate the isotonic requirement; after that, we develop refined procedures for improving

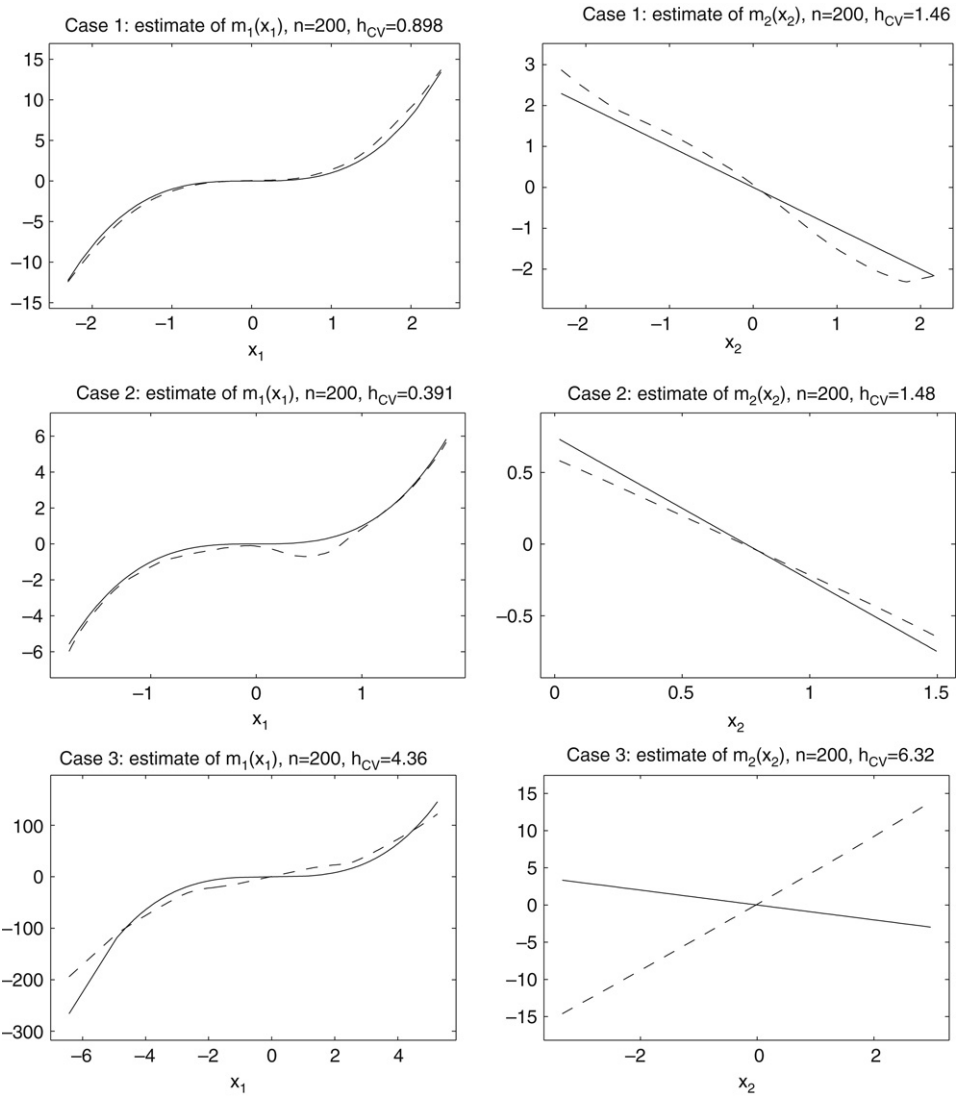

Fig. 1. Component functions m_1 and m_2 in the additive model (4.1). Solid curve: true function; dashed curve: fitted function via local linear backfitting procedure.

the original estimates and ensure that the modified estimates are monotone on its entire domain. An implicit assumption underlying the above developments is the independence between all covariates.

This paper differs from existing results in a number of ways. First, this paper discusses, in particular for multi-dimensional parametric, non-parametric and semi-parametric models, that due to dependence mechanism, some function estimates completely violate the monotone requirement (as observed in the third case of Fig. 1). Thus it is impossible to locally modify or improve the original estimates. Secondly, this paper intends to help provide a better understanding of why this occurs and how to circumvent it. Stein's Lemma ([21] Lemma 2),

which is important in the theory of statistics and probability and in applications to capital asset pricing models ([19, Sec. 4.5]; [2, p. 164]), will be used as a technical tool in our investigation.

The rest of the article is organized as follows. Section 2 extends the ordinary Stein's Lemma for mutually independent Gaussian random variables to jointly dependent Gaussian random variables, and then to independent non-Gaussian random variables. An interesting connection to support vector machine is mentioned there. Section 3 proposes a de-correlation transform which deals with arbitrarily dependent random variables. Section 4 applies the results to commonly used multi-dimensional models. Section 5 analyzes real data, and Section 6 ends the paper with a brief concluding remark. Technical proofs are relegated to the Appendix.

2. Stein's lemma and functional monotonicity

In this section, we demonstrate that under mild conditions, the isotonic directions of a multi-dimensional function can be captured on a numerical scale. For convenience, we first introduce some necessary notation. We consider a d-variate random vector, $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, where the superscript $^{\mathrm{T}}$ denotes transpose. Denote by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^{\mathrm{T}} = E(\boldsymbol{X})$ the mean vector and by $\Gamma = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})$ the covariance matrix. Throughout the paper, we assume the existence and finiteness of $\boldsymbol{\mu}$ and the positive definiteness of Γ. Write $\mathbf{I}$ for an identity matrix. A function $m : \mathbb{R}^d \to \mathbb{R}$ is said to be *almost differentiable* if there exists a function $f : \mathbb{R}^d \to \mathbb{R}^d$ such that, for all $\boldsymbol{z} \in \mathbb{R}^d$, $m(\boldsymbol{x} + \boldsymbol{z}) - m(\boldsymbol{x}) = \int_0^1 \boldsymbol{z}^{\mathrm{T}} f(\boldsymbol{x} + t\boldsymbol{z})\mathrm{d}t$ for almost all $\boldsymbol{x} = (x_1, \ldots, x_d)^{\mathrm{T}} \in \mathbb{R}^d$. Then f is essentially unique; f is called the *gradient* of m and denoted by $\partial m(\boldsymbol{x})/\partial \boldsymbol{x} = (\partial m(\boldsymbol{x})/\partial x_1, \ldots, \partial m(\boldsymbol{x})/\partial x_d)^{\mathrm{T}}$. Define

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}} = \Gamma^{-1} \mathrm{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\}. \tag{2.1}$$

Since a function $m(\boldsymbol{x})$ is not necessarily monotone in all its variables, we consider, throughout the paper, a partially monotone function $m(\boldsymbol{x})$ which is monotone in a subset of its coordinates, say in its first J components, where $1 \le J \le d$ and J is known. This arises naturally from multi-dimensional modeling where the number d of covariates far exceeds the number J of monotone directions. We would anticipate that an estimate of the unknown $m(\boldsymbol{x})$ from noisy data inherits the monotone directions in those J coordinates. To this end, we will first study the linkage between the signs of $\boldsymbol{\theta}$ and the isotonic directions of $m(\boldsymbol{x})$.

The study on the linkage property is motivated from the celebrated Stein's Lemma. For a random vector $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \mathbf{I})$ consisting of mutually independent Gaussian random variables, Stein's Lemma states that

$$E\{(\boldsymbol{X} - \boldsymbol{\mu})m(\boldsymbol{X})\} = E\left\{\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{X}}\right\}, \tag{2.2}$$

for an almost differentiable function $m(\boldsymbol{x})$ with $E\{|\partial m(\boldsymbol{X})/\partial X_j|\} < \infty$, $j = 1, \ldots, d$. In this case, since $\Gamma = \mathbf{I}$ and the left side of (2.2) can be rewritten as $\mathrm{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\}$, we observe the coincidence $\boldsymbol{\theta} = E\{\partial m(\boldsymbol{X})/\partial \boldsymbol{X}\}$. Thus for any index $j \in \{1, \ldots, J\}$, $m(\boldsymbol{x})$ being monotone increasing in x_j implies $\partial m(\boldsymbol{x})/\partial x_j \ge 0$ and therefore $\theta_j \ge 0$; similarly, $\theta_j \le 0$ if $m(\boldsymbol{x})$ is monotone decreasing in x_j. Hence, if the component variables of $\boldsymbol{X}$ have the distribution,

$$\text{Case 0: both jointly Gaussian and mutually independent,} \tag{2.3}$$

then the first J signs of $\boldsymbol{\theta}$ map the J monotone directions of $m(\boldsymbol{x})$, with a positive sign corresponding to the monotone increasing direction and a negative sign the monotone decreasing direction.

Thus, under the assumption (2.3), $\boldsymbol{\theta}$ serves as a surrogate vector for characterizing the monotone directions of $m(\boldsymbol{x})$. This nice linkage property enjoyed by $\boldsymbol{\theta}$ offers technical convenience and practical guidance for studying functional monotonicity. A more careful study is needed to investigate whether assumption (2.3) can be extended to other situations without losing generality.

In realistic applications, covariates may be neither mutually independent nor normally distributed, i.e., assumption (2.3) is too restrictive. To learn whether the linkage property between the surrogate vector and the monotone directions continues to hold for covariates that are either dependent or non-Gaussian, there is a need to generalize Stein's Lemma to incorporate random vectors $\boldsymbol{X}$ whose distribution deviates from (2.3). Three cases below are of primary interest to the generalization and will be discussed in Sections 2.1, 2.2 and 3, respectively.

Case I: jointly Gaussian and mutually dependent (i.e., dropping the independence assumption in (2.3)).

Case II: jointly non-Gaussian and mutually independent (i.e., dropping the normality assumption in (2.3)).

Case III: jointly non-Gaussian and mutually dependent (i.e., dropping both the normality and independence assumptions in (2.3)).

Remark 1. From a statistical point of view, $\boldsymbol{\theta}$ is a vector of parameters. In practice, we do not know its true value, but could estimate the desired quantity. Consider, for example, $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, whose applications are detailed in Section 4. For a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ consisting of i.i.d. observation pairs, where $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{di})^{\mathrm{T}}$, set $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_d)^{\mathrm{T}}$, where $\widehat{\mu}_j = \sum_{i=1}^n X_{ji}/n$, $j = 1, \ldots, d$. By the law of large numbers, a consistent estimator for Γ is $\widehat{\Gamma} = \sum_{i=1}^n (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})^{\mathrm{T}}/(n-1)$, and $\widehat{\mathrm{cov}} = \sum_{i=1}^n (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}})Y_i/(n-1)$ is a consistent estimator of $\mathrm{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\}$. Hence a consistent estimator for $\boldsymbol{\theta}$ in Cases 0, I, II and III can be formed by

$$\widehat{\boldsymbol{\theta}} = \widehat{\Gamma}^{-1}\,\widehat{\mathrm{cov}}.$$

Clearly, $\widehat{\boldsymbol{\theta}}$ is easy to obtain and conveniently facilitates data analysis. Even if $\widehat{\boldsymbol{\theta}}$ is biased, we could rely on signs of $\widehat{\boldsymbol{\theta}}$ to correctly identify the monotone directions, as long as the signs are right. In contrast, the monotonicity of an unknown function $m(\boldsymbol{x})$ is an inherent analytical property, i.e., directly checking the direction of monotonicity from noisy data will be much more challenging and complicated. This approach is similar in spirit to the support vector machine [12], in which the corresponding classification rule is induced only by the sign of some discriminating function, but not by its actual value.

2.1. Case I

For a trivariate normal random vector, a three-dimensional version of Stein's Lemma can be found in [7]. Here, the present paper concerns the linkage property for distributions of $\boldsymbol{X}$ belonging to Case I. An example below lends numerical support for the desired linkage property.

Example 1. Consider a bivariate normal random vector $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}}$ with mean vector zero and covariance matrix given by

$$\Gamma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}, \tag{2.4}$$

where $|\rho| < 1$. For a function $m(x_1, x_2) = x_1^3 - x_2$, which is monotone increasing in x_1 and decreasing in x_2, the signs of $\boldsymbol{\theta}$ defined in (2.1) give the correct monotone directions in x_1 and x_2, i.e., $\theta_1 > 0$ and $\theta_2 < 0$. To verify this, an explicit evaluation of $\boldsymbol{\theta}$ can be obtained from the derivations,

$$\Gamma^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix},$$

$\text{cov}\{X_1, m(\boldsymbol{X})\} = 3 - \rho$, and $\text{cov}\{X_2, m(\boldsymbol{X})\} = E(X_1^3 X_2) - 1$, in which $E(X_1^3 X_2) = 3\rho$, thus

$$\text{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\} = \begin{bmatrix} 3-\rho \\ 3\rho - 1 \end{bmatrix}.$$

By the definition of $\boldsymbol{\theta}$ in (2.1),

$$\boldsymbol{\theta} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\begin{bmatrix} 3-\rho \\ 3\rho - 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}.$$

Theorem 1 formalizes the assertion that in Case I there is an explicit one-to-one correspondence between the jth monotone direction of a multi-dimensional function and the jth sign of the surrogate vector. For notational simplicity, we denote by $\boldsymbol{X}_{-j}$ the part of $\boldsymbol{X}$ excluding X_j.

Theorem 1 (*Case I*). *For an index $j \in \{1, \ldots, J\}$, suppose that X_j and $\boldsymbol{X}_{-j}$ have a jointly Gaussian distribution. For an almost differentiable function $m : \mathbb{R}^d \to \mathbb{R}$, assume that for each $k = 1, \ldots, d$, $\partial m(\boldsymbol{x})/\partial x_k$ exists almost everywhere fulfilling $E\{|\partial m(\boldsymbol{X})/\partial X_k|\} < \infty$. We have that*

(i) *If $m(\boldsymbol{x})$ is monotone increasing in the jth coordinate x_j, then $\theta_j \geq 0$; moreover, in this case, $\theta_j > 0$ if and only if $P\{\partial m(\boldsymbol{X})/\partial X_j > 0\} > 0$.*
(ii) *If $m(\boldsymbol{x})$ is monotone decreasing in the jth coordinate x_j, then $\theta_j \leq 0$; moreover, in this case, $\theta_j < 0$ if and only if $P\{\partial m(\boldsymbol{X})/\partial X_j < 0\} > 0$.*

2.2. Case II

Regarding Case II, Theorem 2 below demonstrates that the assumption of jointly Gaussian distribution in Theorem 1 can indeed be replaced by independence. For $j = 1, \ldots, d$, suppose that X_j takes values in the domain Ω_j.

Theorem 2 (*Case II*). *For an index $j \in \{1, \ldots, J\}$, suppose that X_j is independent of $\boldsymbol{X}_{-j}$. Assume that $m : \Omega_1 \times \cdots \times \Omega_d \to \mathbb{R}$ is a measurable function. Assume that $E\{|m(\boldsymbol{X})|\} < \infty$ and $E\{|X_j m(\boldsymbol{X})|\} < \infty$. Define $B_j = E\{(X_j - \mu_j)m(\boldsymbol{X})|\boldsymbol{X}_{-j}\}$. We have that*

(i) *If $m(\boldsymbol{x})$ is monotone increasing in the jth coordinate x_j, then $\theta_j \geq 0$; moreover, in this case, $\theta_j > 0$ if and only if $P(B_j > 0) > 0$.*
(ii) *If $m(\boldsymbol{x})$ is monotone decreasing in the jth coordinate x_j, then $\theta_j \leq 0$; moreover, in this case, $\theta_j < 0$ if and only if $P(B_j < 0) > 0$.*

Remark 2. For a univariate ($d = 1$) random variable X with expectation $\mu = E(X)$, following the Tchebychef's inequality ([10], p. 43 and 168) used in the proof of Theorem 2, we observe that if $m(x)$ is monotone increasing in x, then $\theta \geq 0$ (namely, $\text{cov}\{X, m(X)\} \geq 0$); moreover,

in this case, $\theta > 0$ (namely, $\mathrm{cov}\{X, m(X)\} > 0$) if and only if $P\{(X-\mu)m(X) > 0\} > 0$. Analogous results definitely hold for monotone decreasing functions. As such, the utility of the surrogate vector $\boldsymbol{\theta}$ is particularly reflected in applications to larger-dimensional models with two or more explanatory variables.

A number of insights can be obtained from a comparison of Theorems 1 and 2. First, the proof of Theorem 1 depends on the conventional form (2.2) of Stein's Lemma, whereas the proof of Theorem 2 does not use Stein's Lemma. Second, in terms of the distributional assumption, Theorem 1 requires joint normality of X_j and $\boldsymbol{X}_{-j}$, whereas Theorem 2 demands independence of X_j and $\boldsymbol{X}_{-j}$, but does not require the existence of either their joint probability density function or their marginal probability density functions. This relaxation is especially useful for applications to non-Gaussian input variables, such as Bernoulli and Poisson variables. Third, with respect to the smoothness assumption, Theorem 1 constrains the function $m(\boldsymbol{x})$ to be almost differentiable in all its arguments, even if $m(\boldsymbol{x})$ is monotone in some of them but is not monotone in the remaining arguments. It should be stressed that in some applications verifying the almost differentiability of an unknown m is not easy. Indeed, this strong assumption is removed from Theorem 2. Actually, by standard analysis arguments [20, p. 96], if $m(\boldsymbol{x})$ is monotone in $x_j \in \Omega_j$, then $m(\boldsymbol{x})$, when viewed as a univariate function of x_j, can only have countably many discontinuities (of jump type) in x_j. Thus Theorem 2 avoids an explicit smoothness assumption on $m(\boldsymbol{x})$ with respect to the remaining arguments.

2.3. Case III

The most challenging case is Case III. The comparison in the preceding paragraph indicates that Case II is fully adaptive to non-Gaussian distributions of covariates. This relaxation inspires us to explore an extension of Theorem 2 to Case III. However, it is worth mentioning that the proof of Theorem 2 depends critically on the validity of (A.8) (in the Appendix). Namely, (A.8) may not be ensured if the independence assumption is substituted by an arbitrary structure of dependence. To elaborate this point, Example 2 below illustrates the potential effects of dependent covariates on the breakdown of (A.8). Hence, for covariates that are dependent, the signs of the surrogate vector may fail to reveal the monotone directions. A more thorough discussion of Case III will be given in Section 3.

Example 2. Consider dependent covariates $(X_1, X_2)^{\mathrm{T}}$ which follow a mixture of bivariate normal distributions $N\{(a_1, 1)^{\mathrm{T}}, \mathbf{I}\}$ and $N\{-(a_1, 1)^{\mathrm{T}}, \mathbf{I}\}$ with equal mixing proportions. Apparently, this joint distribution does not fall in either Case I or Case II. For a function $m(x_1, x_2) = \exp(-x_1) - x_2$, which decreases in both x_1 and x_2, it can be shown that for $j = 1$, the left side of (A.8) is

$$B_1 = \frac{\{\mathrm{e}^{1/2-a_1}(a_1-1) - a_1X_2\}\phi(X_2-1) - \{\mathrm{e}^{1/2+a_1}(a_1+1) - a_1X_2\}\phi(X_2+1)}{\phi(X_2-1)+\phi(X_2+1)},$$

whereas the right side of (A.8) equals

$$2^{-1}\{\mathrm{e}^{1/2-a_1}(a_1-1) - \mathrm{e}^{1/2+a_1}(a_1+1)\}.$$

Hence, (A.8) does not hold.

As a consequence, the explicit evaluation of $\boldsymbol{\theta}$ follows directly from

$$\Gamma = \begin{bmatrix} a_1^2+1 & a_1 \\ a_1 & 2 \end{bmatrix} \text{ and } \mathrm{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\} = \begin{bmatrix} \{\mathrm{e}^{1/2-a_1}(a_1-1) - \mathrm{e}^{1/2+a_1}(a_1+1)\}/2 - a_1 \\ \{\mathrm{e}^{1/2-a_1} - \mathrm{e}^{1/2+a_1}\}/2 - 2 \end{bmatrix}.$$

For $a_1 = 3$, it is readily seen that $\theta_1 < 0$ but $\theta_2 > 0$. This indicates that the sign of θ_2 does not correctly mirror the monotone decreasing direction of $m(\boldsymbol{x})$ in x_2.

Interestingly, Stein's Lemma has recently been generalized to the class of Elliptical distributions, which includes the multivariate Gaussian, Student-t, Cauchy, symmetric stable and many other distributions (see [6]). In [13,14], the generalization of Stein's Lemma for bivariate and multivariate Elliptical distributions are discussed, respectively. It remains unclear whether, and under which assumptions, the linkage property can be extended to the broader class of Elliptical distributions.

3. De-correlation transform for dependent random variables

In Case III, there are infinitely many varieties of dependence mechanism between covariates. Though we do not intend to exhaust all possibilities, we will focus on cases of more direct relevance to practical applications. Emphasis will be sequentially put on unspecified dependence, un-correlation, and de-correlation.

In the meanwhile, for modeling and interpretation purposes, the structure of a multi-dimensional function $m(\boldsymbol{x})$ should be as flexible as possible, but not excessively arbitrary. Interestingly, Theorem 3 shows that when $m(\boldsymbol{x})$ is linearly associated with $\boldsymbol{x}$, signs of the surrogate vector $\boldsymbol{\theta}$ continue to reveal the monotone directions, regardless of the dependence mechanism of $\boldsymbol{X}$.

Theorem 3 (*Unspecified Dependence*). *Suppose that the covariance matrix Γ of a random vector $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ exists and is positive definite. If $m(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}$ with parameters β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$, then $\boldsymbol{\theta} \equiv \boldsymbol{\beta}$. Thus the signs of $\boldsymbol{\theta}$ invariably accord with the monotone directions of $m(\boldsymbol{x})$.*

Theorems 4 and 5 below will remove the linearity relationship in Theorem 3. In particular, Theorem 4 permits the partially non-linear association, whereas Theorem 5 further allows the non-linear association in an additive manner.

Theorem 4 (*Un-correlation*). *For an index $j \in \{1, \ldots, J\}$, assume the conditions*

(C1) *X_j is un-correlated with $\boldsymbol{X}_{-j}$;*

(C2) *$m : \Omega_1 \times \cdots \times \Omega_d \to \mathbb{R}$ is a measurable function, and $m(\boldsymbol{x})$ can be represented in the form, $m(\boldsymbol{x}) = m_j(x_j) + \boldsymbol{\beta}_{-j}^{\mathrm{T}}\boldsymbol{x}_{-j}$, in which the function $m_j(\cdot)$ is unspecified, $\boldsymbol{\beta}_{-j}$ is a vector of $d-1$ parameters and $\boldsymbol{x}_{-j}$ is the part of $\boldsymbol{x}$ excluding x_j.*

Also, suppose $E\{|m_j(X_j)|\} < \infty$ and $E\{|X_j m_j(X_j)|\} < \infty$. We have that

(i) *If $m(\boldsymbol{x})$ is monotone increasing in the jth coordinate x_j, then $\theta_j \geq 0$; moreover, in this case, $\theta_j > 0$ if and only if $P\{(X_j - \mu_j)m_j(X_j) > 0\} > 0$.*

(ii) *If $m(\boldsymbol{x})$ is monotone decreasing in the jth coordinate x_j, then $\theta_j \leq 0$; moreover, in this case, $\theta_j < 0$ if and only if $P\{(X_j - \mu_j)m_j(X_j) < 0\} > 0$.*

Before proceeding with Theorem 5, we will need the following definition.

Definition 1. A random variable X_1 is called "de-correlated" with a $\ell \times 1$ random vector $\boldsymbol{X}_2$ if X_1 is un-correlated with any $f(\boldsymbol{X}_2)$, where $f : \mathbb{R}^{\ell} \to \mathbb{R}$ is a measurable function.

Table 1
Summary of Theorems 3 up to 5 in Case III with $m(\boldsymbol{x}) = m_j(x_j) + M_j(\boldsymbol{x}_{-j})$

$m_j(x_j)$	$M_j(\boldsymbol{x}_{-j})$	
	Linear	Unspecified
linear	Theorem 3: arbitrary dependence of $\boldsymbol{X}$	Theorem 5: X_j de-correlated with $\boldsymbol{X}_{-j}$
unspecified	Theorem 4: X_j un-correlated with $\boldsymbol{X}_{-j}$	Theorem 5: X_j de-correlated with $\boldsymbol{X}_{-j}$

The condition of "de-correlated" is stronger than that of "un-correlated" assumed in Theorem 4, but indeed weakens the assumption of independence in Theorem 2. Based on this notion, Theorem 5 offers an extension of Theorem 2. It can be seen that major technical arguments used in Theorem 2 go through to those in Theorem 5.

Theorem 5 (*De-correlation*). *For an index $j \in \{1, \ldots, J\}$, assume the conditions*

(D1) *X_j is de-correlated with $\boldsymbol{X}_{-j}$;*
(D2) *$m : \Omega_1 \times \cdots \times \Omega_d \to \mathbb{R}$ is a measurable function, and $m(\boldsymbol{x})$ can be represented in the form, $m(\boldsymbol{x}) = m_j(x_j) + M_j(\boldsymbol{x}_{-j})$, in which both functions $m_j(\cdot)$ and $M_j(\cdot)$ are unspecified and $\boldsymbol{x}_{-j}$ is the part of $\boldsymbol{x}$ excluding x_j.*

Also, suppose $E\{|m_j(X_j)|\} < \infty$ and $E\{|X_j m_j(X_j)|\} < \infty$. We have that

(i) *If $m(\boldsymbol{x})$ is monotone increasing in the jth coordinate x_j, then $\theta_j \geq 0$; moreover, in this case, $\theta_j > 0$ if and only if $P\{(X_j - \mu_j)m_j(X_j) > 0\} > 0$.*
(ii) *If $m(\boldsymbol{x})$ is monotone decreasing in the jth coordinate x_j, then $\theta_j \leq 0$; moreover, in this case, $\theta_j < 0$ if and only if $P\{(X_j - \mu_j)m_j(X_j) < 0\} > 0$.*

Is assumption (D1) purely for the ease of technical proofs or practically unrealistic to be achieved? To answer this question, we notice the fact that for any random variable X_1 and any random vector $\boldsymbol{X}_2$, if $E(|X_1|) < \infty$, then a transformed variable defined by

$$X_{1|2} = X_1 - E(X_1|\boldsymbol{X}_2),$$

is un-correlated with not only any linear function of $\boldsymbol{X}_2$ but also any measurable function of $\boldsymbol{X}_2$. Moreover, if X_2 is univariate, then the projection part, $E(X_1|X_2 = x_2)$, can easily be estimated by a one-dimensional nonparametric regression technique, such as smoothing splines, regression splines, and the local polynomial regression method. Likewise, the smoothing parameter can simply be chosen to minimize the cross-validation criterion. Thus, the "de-correlation" procedure is applicable to achieve both assumption (C1) in Theorem 4 and assumption (D1) in Theorem 5.

Assumption (D2) in Theorem 5 is also very broad, including modeling functions arising from the additive regression model, semi-parametric partially linear model, generalized linear model, and many others. These models relax the stringent assumption of linearity in Theorem 3, which restricts applications to the linear regression model, and they will be addressed in detail in the next section.

Before ending this section, Table 1 summarizes the domains of applications of Theorems 3–5.

4. Applications to estimating monotone functions in multi-dimensional models

Theorems 1–5 of the preceding Sections 2 and 3 discuss the relation between the signs of $\boldsymbol{\theta}$ and the monotone directions of $m(\boldsymbol{x})$. In practice, $m(\boldsymbol{x})$ is unknown and needs to be estimated.

Table 2
Percentage of samples from the additive model (4.1) with correct direction of monotonicity

Sample size	Case 1	Case 2	Case 3
200	96.75%	100%	25.50%
400	98.25%	100%	26.75%

Let $\widehat{m}(\boldsymbol{x})$ denote a data-based estimate of $m(\boldsymbol{x})$, using a random sample $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$. In this section, we will apply the general results of Theorems 1–5 to $m(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, the mean regression function of a response variable Y on covariates $\boldsymbol{X}$, and will investigate the discrepancy between the observed monotone directions of $\widehat{m}(\boldsymbol{x})$ and the actual monotone directions of $m(\boldsymbol{x})$.

4.1. Nonparametric estimation of monotone functions in additive regression model

Estimating a multi-dimensional regression function is a challenging task. To overcome the "curse of dimensionality", additive modeling has been proposed as an efficient technique [11]. An additive regression model assumes that $Y = m(\boldsymbol{X}) + \varepsilon$, where $E(\varepsilon|\boldsymbol{X} = \boldsymbol{x}) = 0$, and the regression function is a sum of smooth functions of component variables, i.e.,

$$m(\boldsymbol{x}) = \alpha + m_1(x_1) + \cdots + m_d(x_d),$$

for a parameter α and univariate functions $m_1, \ldots, m_d$. To ensure identifiability, the conditions $E\{m_j(X_j)\} = 0$, $j = 1, \ldots, d$, are usually imposed. Since the change in $m(\boldsymbol{x})$ with respect to x_j is precisely the change in $m_j(x_j)$ with respect to x_j, the monotone direction of $m(\boldsymbol{x})$ in x_j is exactly the same as that of the jth component function $m_j(x_j)$ in x_j.

As an illustration, we revisit Fig. 1 in Section 1. Random samples of (X_1, X_2, Y) are generated according to a bivariate additive model,

$$Y = \alpha + m_1(X_1) + m_2(X_2) + \varepsilon, \tag{4.1}$$

with $\alpha = 0$, $m_1(x_1) = x_1^3 - E(X_1^3)$ and $m_2(x_2) = -\{x_2 - E(X_2)\}$. Clearly, $m(\boldsymbol{x})$ has two monotone directions, increasing in x_1 and decreasing in x_2. The error $\varepsilon \sim N(0, \sigma^2)$ is independent of X_1 and X_2. The distributions of X_1 and X_2 are examined in three cases, and in each case the magnitude of σ is chosen so that the signal-to-noise ratio (SNR) is about 5. Here SNR is defined as $\mathrm{var}\{E(Y|\boldsymbol{X})\}/E\{\mathrm{var}(Y|\boldsymbol{X})\}$.

Case 1: $(X_1, X_2)^{\mathrm{T}}$ follows a bivariate normal distribution $N(\mathbf{0}, \Gamma)$, with Γ given by (2.4) where $\rho = .7; \sigma = 1.5$.

Case 2: X_1 and X_2 are independent where $X_1 \sim \mathrm{Uniform}(-1.8, 1.8)$ and $X_2 \sim \mathrm{Uniform}(0, 1.5)$; $\sigma = 1$.

Case 3: $(X_1, X_2)^{\mathrm{T}}$ follows a mixture of bivariate normal distributions $N\{(a_1, 1)^{\mathrm{T}}, \mathbf{I}\}$ and $N\{-(a_1, 1)^{\mathrm{T}}, \mathbf{I}\}$ with equal mixing proportions and $a_1 = -3$; $\sigma = 22$.

Fig. 1 presents the fitted curves via local-linear backfitting procedure. The data-driven choice of bandwidth in the backfitting iterations is selected by the cross-validation criterion. Throughout the paper, the Epanechnikov kernel is used. In Cases 1–2, direct calculations give $\theta_1 > 0$ and $\theta_2 < 0$. Theorems 1 and 2 guarantee that the monotone directions of $m(\boldsymbol{x})$ agree with the signs of $\boldsymbol{\theta}$. The fitted curves indeed follow the correct monotone directions. In Case 3, explicit calculations give $\theta_1 > 0$ but $\theta_2 > 0$. However, we do not have theoretical results concluding that the monotone directions of $m(\boldsymbol{x})$ concur with the signs of $\boldsymbol{\theta}$. Actually, the fitted curve for $m_1(x_1)$ produces a correct monotone direction in x_1, but the fitted curve for $m_2(x_2)$ produces an

incorrect monotone direction in x_2. To assess whether the observed discrepancy is likely to be due to chance fluctuations, we repeat the simulations for 400 times. Only if both fitted curves in one simulation correctly reveal the monotone directions did we report a correct direction of monotonicity. The percentage of samples with correct direction of monotonicity is given in Table 2. Table 2 lends further support that for $\boldsymbol{X}$ with dependent covariates such as in Case 3, the nonparametric function estimates for the additive model, which is commonly used in practice, may produce incorrect monotone directions, with a non-ignorable high probability (around 75%). This undesirable result could hardly be eased with an increase in sample size (from 200 to 400), in which case the performance of function estimates will enhance. This in turn indicates that the discrepancy between the actual monotone directions and the observed monotone directions is mainly caused by the dependence of covariates, and is negligibly influenced by the bias of nonparametric function estimates.

The phenomenon observed from Case 3 is not unique to the example given there. Similar examples could be constructed. For instance, consider $X_1 \sim \text{Uniform}(-1, 1)$, $X_2 = X_1^2$, $m_1(x_1) = \exp(-x_1) - E\{\exp(-X_1)\}$, $m_2(x_2) = -.1\{x_2^3 - E(X_2^3)\}$, and $\sigma = .3$. Clearly, $m(\boldsymbol{x})$ is monotone decreasing in both x_1 and x_2. The local linear backfitting produces a correct monotone direction of $m(\boldsymbol{x})$ in x_1, but an incorrect monotone direction in x_2. This also coincides with the explicit calculations $\theta_1 < 0$ and $\theta_2 > 0$. For constructions of other types of dependent covariates, we refer to [4,15] for many interesting and insightful results on dependence structures of random variables.

4.2. Parametric estimation in semi-parametric partially linear model

In the semi-parametric partially linear model with the covariate vector $\boldsymbol{X} = (U, Z_1, \ldots, Z_q)^{\mathrm{T}}$, the regression function at $(u, \boldsymbol{z})$ takes the form,

$$m(u, \boldsymbol{z}) = a(u) + \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\beta},$$

where $u \in \mathbb{R}$, $\boldsymbol{z} = (z_1, \ldots, z_q)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ is a vector of unknown parameters, and the unknown smooth function $a(u)$ nonparametrically describes the effect of U on the mean response. For detailed information, we refer in particular to [9]. Often, interest centers on the statistical estimation and inference for the parametric component $\boldsymbol{\beta}$. It is easy to see that $\partial m(u, \boldsymbol{z})/\partial u = a'(u)$ if $a(u)$ is differentiable, and that $\partial m(u, \boldsymbol{z})/\partial z_j = \beta_j$, $j = 1, \ldots, q$. Thus $m(u, \boldsymbol{z})$ is monotone in each of the arguments $z_1, \ldots, z_q$, with directions of monotonicity completely determined by the signs of $\boldsymbol{\beta}$.

For the purpose of illustration, we consider a partially linear model

$$Y = a(U) + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon, \tag{4.2}$$

consisting of three covariates, where $a(u) = u^3$, $\beta_1 = -1$, $\beta_2 = 1$, (U, Z_1) follows the joint distribution specified in three cases of Section 4.1, and $Z_2 \sim \text{Uniform}(0, 1)$ is independent of (U, Z_1). The error $\varepsilon \sim N(0, \sigma^2)$ is independent of (U, Z_1, Z_2), with the same choice of σ as specified in Section 4.1. Fig. 2 displays the local linear estimate of $a(u)$ and the profile least-squares estimate of $\boldsymbol{\beta}$. The selection of bandwidth used in estimating $a(u)$ applies the method in [23]. In Cases 1–2, the signs of the estimates of $\boldsymbol{\beta}$ are correct. Nonetheless, in Case 3, we observe incorrect signs of the estimates of $\boldsymbol{\beta}$. We run the simulation 400 times, and Table 3 summarizes the proportions of times both the fitted curves and parametric estimates correctly reveal the monotone directions with respect to U, Z_1 and Z_2. The implication further supports our analysis in Section 4.1.

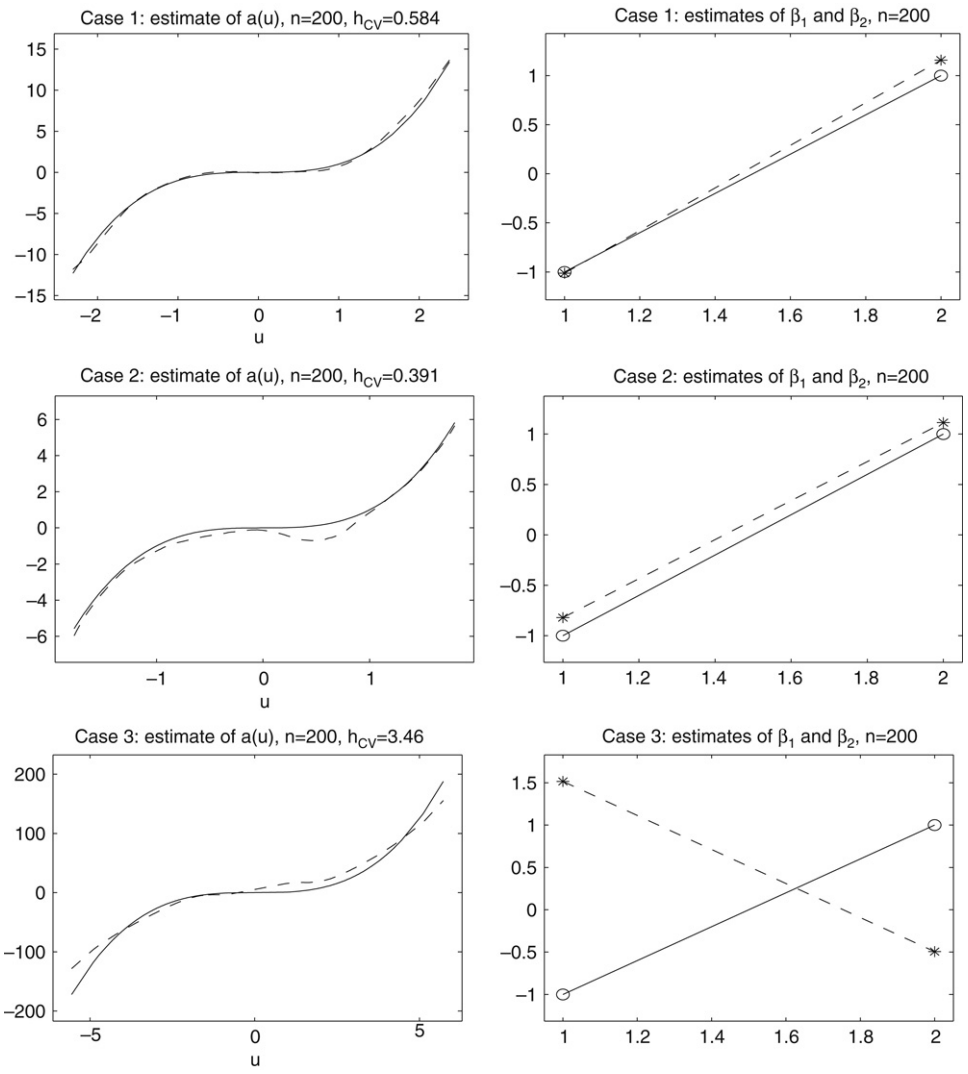

Fig. 2. Nonparametric and parametric components in the partially linear model (4.2). In the left column, solid curves denote true functions and dashed curves denote the fitted functions. In the right column, circles denote the true values of β_1 and β_2 (connected by solid lines), and stars denote the estimated values of β_1 and β_2 (connected by dashed lines).

Table 3
Percentage of samples from the partially linear model (4.2) with correct direction of monotonicity

Sample size	Case 1	Case 2	Case 3
200	99.5%	100%	41.75%
400	100%	100%	53.25%

4.3. Parametric estimation in generalized linear model

The generalized linear model (GLM) is commonly used for the response variable Y which, conditional on the covariate vector $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, has a distribution in the exponential

Table 4
Percentage of samples from the generalized linear model (4.4) with correct direction of monotonicity

Response variable	Sample size	Case 1	Case 2	Case 3
Gaussian	200	100%	100%	43.5%
	400	100%	100%	54.5%

family, taking the form

$$f_{Y|\boldsymbol{X}}(y;\theta(\boldsymbol{x})) = \exp[\{y\theta(\boldsymbol{x}) - b(\theta(\boldsymbol{x}))\}/a(\psi) + c(y,\psi)],$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$, where $\theta(\boldsymbol{x})$ is called a canonical parameter and ψ is called a dispersion parameter, respectively. It is well known that $m(\boldsymbol{x}) \equiv E(Y|\boldsymbol{X}=\boldsymbol{x}) = b'(\theta(\boldsymbol{x}))$ and $\sigma^2(\boldsymbol{x}) \equiv \mathrm{var}(Y|\boldsymbol{X}=\boldsymbol{x}) = a(\psi)b''(\theta(\boldsymbol{x}))$. See [18,16]. The GLM assumes that the transformation of the regression function, via a link function g, can be linearly modeled by

$$g(m(\boldsymbol{x})) = \beta_0 + \beta_1 h_1(x_1) + \cdots + \beta_d h_d(x_d), \tag{4.3}$$

for unknown parameters $(\beta_0, \beta_1, \ldots, \beta_d)$ and known monotone functions h_j, $j = 1, \ldots, d$. When g is invertible, (4.3) is equivalent to $m(\boldsymbol{x}) = g^{-1}(\beta_0 + \sum_{j=1}^{d} \beta_j h_j(x_j))$. For any monotonic differentiable link function g, it can be shown that

$$\frac{\partial m(\boldsymbol{x})}{\partial x_j} = \frac{\beta_j h_j'(x_j)}{g'(m(\boldsymbol{x}))}, \quad j = 1, \ldots, d,$$

and hence the monotone direction of $m(\boldsymbol{x})$ in x_j is exclusively determined by the sign of the coefficient β_j. For instance, if $b'(\cdot)$ is invertible and $b''(\cdot)$ is bounded away from 0 and ∞, then for the routinely used canonical link, $g(\cdot) = (b')^{-1}(\cdot)$ (resulting in $g(m(\boldsymbol{x})) = \theta(\boldsymbol{x})$), it follows that

$$g'(m(\boldsymbol{x})) = \frac{1}{b''(\theta(\boldsymbol{x}))} > 0.$$

Examples of the canonical links include $g(m) = m$, $g(m) = \ln\{m/(1-m)\}$ and $g(m) = \ln(m)$ for Gaussian, Bernoulli and Poisson responses, respectively, and in all cases we see that $g'(m) > 0$.

To see whether the monotone directions of the estimates $\widehat{m}(\boldsymbol{x})$ in the GLM are correct, we first generate 400 sets of random samples from a Gaussian regression model,

$$Y = \beta_0 + \beta_1 X_1^3 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon, \tag{4.4}$$

for Gaussian responses. The variables $(X_1, X_2, X_3, \varepsilon)$ have the joint distribution identical to that of $(U, Z_1, Z_2, \varepsilon)$ specified in the three cases of Section 4.2, and the true regression parameters are set to be $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 1, -1, 1)$. Table 4 summarizes the proportions of times (among 400 replicated simulations) the signs of the estimated values of $(\beta_1, \beta_2, \beta_3)$ completely match the signs of the true values of $(\beta_1, \beta_2, \beta_3)$. The non-monotonicity phenomenon continues to occur in Case 3 associated with the generalized linear model.

Next, we conduct a similar simulation study for Bernoulli responses generated from a logistic regression model,

$$\ln\left\{\frac{m(\boldsymbol{X})}{1-m(\boldsymbol{X})}\right\} = \beta_0 + \beta_1 \exp(-X_1) + \beta_2 X_2^3 + \beta_3 X_3,$$

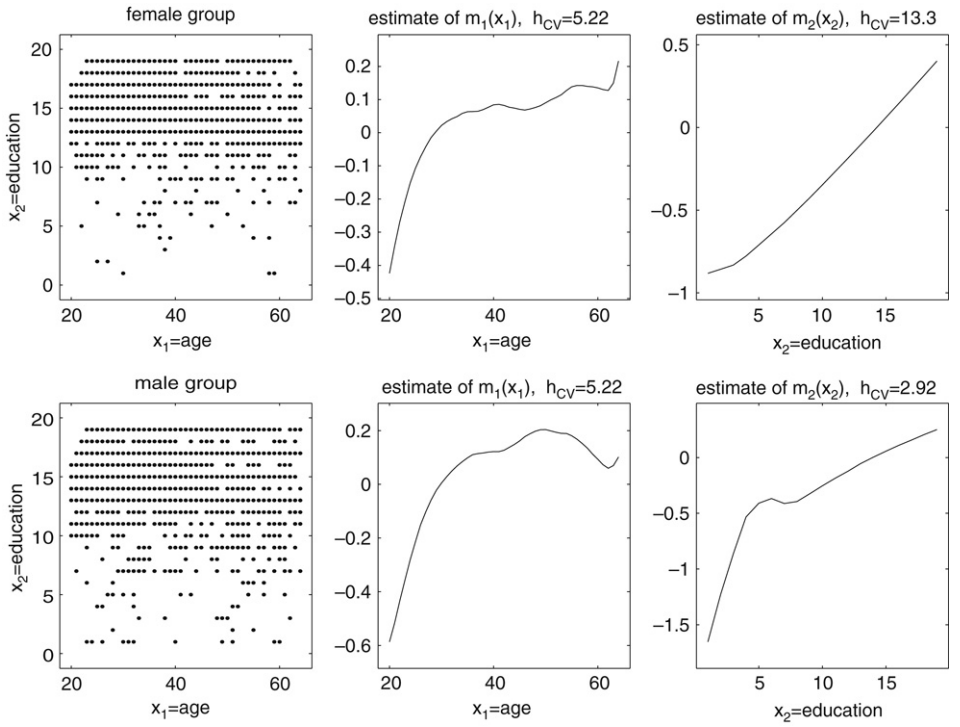

Fig. 3. Additive regression model for ln(wage) on age and education. The top panel corresponds to the female group, whereas the bottom panel corresponds to the male group.

with true parameters $(\beta_0, \beta_1, \beta_2, \beta_3) = (-3.2679, 1, -.1, 3.5496)$, two dependent covariates $X_1 \sim \text{Uniform}(-1, 1)$ and $X_2 = X_1^2$. The third covariate $X_3 \sim \text{Uniform}(0, 1)$ is independent of (X_1, X_2). This example is designed in a similar fashion to the one discussed at the end of Section 4.1. Among 400 replicated simulations, the proportions of times the signs of the estimated values of $(\beta_1, \beta_2, \beta_3)$ completely match the signs of the true values of $(\beta_1, \beta_2, \beta_3)$ are .5975 for $n = 200$, and .5800 for $n = 400$, respectively. The non-monotonicity phenomenon continues to be evident.

5. Real data application

5.1. Example 1

We consider the data set studied in [17]. The data consist of 2447 observations on three variables, ln (wage), age and education, for women. Of interest is to learn how wages vary with years of age and years of education. The scatter plot in the top left corner of Fig. 3 clearly indicates that variables $X_1 =$ age and $X_2 =$ education can practically be treated as independent. Although the distributions of X_1 and X_2 are unknown and deviate significantly from normality, the signs of the estimator $\widehat{\boldsymbol{\theta}}$ associated with $\boldsymbol{X} = (X_1, X_2)^{\mathrm{T}}$ reveal that $\widehat{\theta}_1 > 0$ and $\widehat{\theta}_2 > 0$. According to Theorem 2, we would expect to find an isotonic increasing regression function of $Y =$ ln(wage) in X_1 as well as in X_2. Furthermore, to better distinguish the separate effects of covariates, we fit an additive model for regressing Y on X_1 and X_2. The fitted

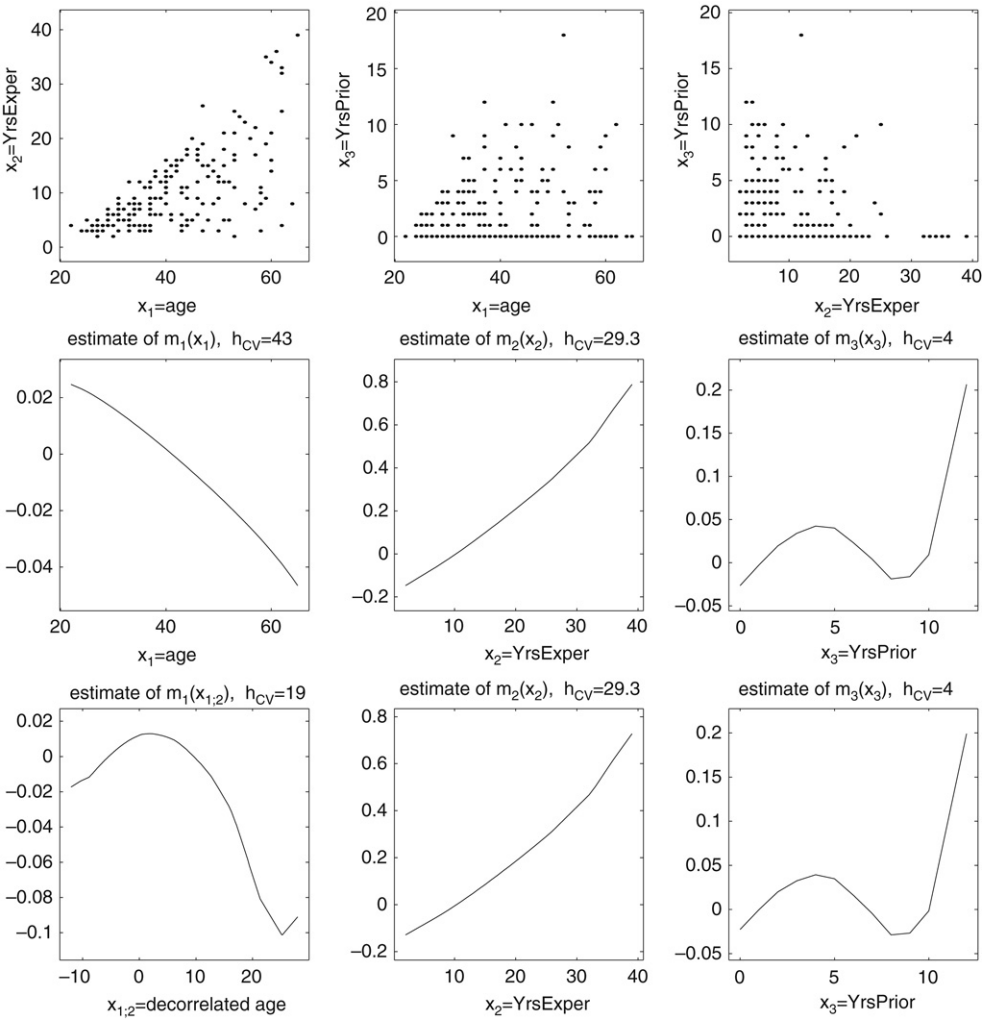

Fig. 4. Additive regression model for the bank salary data. Top panel: pairwise scatter plots of X_1, X_2 and X_3. Middle panel: nonparametric estimates of the component functions. Bottom panel: nonparametric estimates of the component functions, with X_1 replaced by the de-correlated variable $X_{1|2}$.

component functions via local-linear backfitting procedure indeed exhibit the overall upward trend in age and education. The observed violation of monotonicity mainly occurs at boundary sections, where the "boundary bias" problem due to sparseness of data points is well known in nonparametric regression estimation. Similar plots for men are shown in the bottom panel of Fig. 3. Our evaluation, based on the signs of $\widehat{\boldsymbol{\theta}}$ and nonparametric fits, lends support to the predicted result in theoretical and empirical literature in socio-economical studies.

5.2. Example 2

We analyze an employee dataset (Example 11.3 of [1]) of the Fifth National Bank of Springfield, based on year 1995 data. For each of its 208 employees, the dataset consists of

eight variables, including

- YrHired: year that an employee was hired;
- YrBorn: year that an employee was born;
- YrsPrior: years of work experience at another bank before working at the Fifth National bank;
- Salary: current annual salary in thousands of dollars.

To explain variation in salary, we fit an additive model for $Y = \ln(\text{Salary})$ on $X_1 = \text{Age}$, $X_2 = \text{YrsExper}$ and $X_3 = \text{YrsPrior}$, where YrsExper is years of experience with Fifth National Bank and is calculated as 95 minus YrHired. The top panel of Fig. 4 depicts pairwise scatter plots of the three explanatory variables. The variables X_1 and X_2 appear to be highly linearly correlated. The observation with YrsPrior greater than 15 is deleted in our data analysis. The central panel of Fig. 4 displays the fitted component functions via local-linear backfitting procedure. The fitted component curve of Age presents a monotone descending pattern, which would seem to be a most unlikely attribute of the salary/age relation. This phenomenon is accounted for by the dependence between X_1 and X_2. The bottom panel gives the nonparametric estimates of the component functions in the additive model, with X_1 replaced by the de-correlated variable $X_{1|2}$ introduced in Section 3. Clearly, the profile of the estimate of m_1 removes the unexpected decreasing trend, and the de-correlation transform has negligible impact on the estimates of m_2 and m_3.

6. Concluding remarks

In many fields of applications, monotone association in the functional mapping of effects of partial covariates on the response variable needs to be preserved in building multi-dimensional models.

In this paper, we introduce a surrogate vector $\boldsymbol{\theta}$ and show that in Case I and Case II, the signs of $\boldsymbol{\theta}$ regulate the actual directions of monotonicity. For practical purposes, $\boldsymbol{\theta}$ can be consistently estimated by $\widehat{\boldsymbol{\theta}}$ and the signs of $\widehat{\boldsymbol{\theta}}$ facilitates our understanding of whether a proposed multi-dimensional model well reflects the expected directions of monotonicity.

To incorporate covariates in Case III that are jointly non-Gaussian and mutually dependent, we further examine three types of multi-dimensional models which are representative and also easily interpretable. We show that under relaxed assumptions on dependence, the above linkage property between $\boldsymbol{\theta}$ and monotone directions continue to hold. Furthermore, a de-correlation transform is proposed to achieve the relaxed assumption on dependence. The resulting procedure enhances modeling interpretability, with little loss of modeling efficiency.

Acknowledgment

The research is supported in part by National Science Foundation grants DMS-0353941, DMS-0705209 and Wisconsin Alumni Research Foundation. The authors are grateful to an anonymous referee, the Associate Editor and the Editor for their insightful comments and suggestions which greatly improve the presentation of the paper.

Appendix. Proofs of main results

A.1. Proof of Theorem 1

From the assumption, we write $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Gamma)$. Using the equivalent representation $\boldsymbol{X} = \boldsymbol{\mu} + \Gamma^{1/2}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N(\mathbf{0}, \mathbf{I})$, it follows that

$$\begin{aligned}\operatorname{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\} &= \operatorname{cov}\{\boldsymbol{\mu} + \Gamma^{1/2}\boldsymbol{Z}, m(\boldsymbol{X})\} \\ &= \Gamma^{1/2}\operatorname{cov}\{\boldsymbol{Z}, m(\boldsymbol{X})\}. \end{aligned} \tag{A.1}$$

For a standard multivariate normal random vector $\boldsymbol{Z}$, an application of Stein's Lemma ([21, Lemma 2]; also see (2.2)) implies that

$$\operatorname{cov}\{\boldsymbol{Z}, m(\boldsymbol{X})\} = E\left\{\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{Z}}\right\}. \tag{A.2}$$

For the right side of (A.2), we further deduce that

$$E\left\{\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{Z}}\right\} = E\left\{\frac{\partial \boldsymbol{X}}{\partial \boldsymbol{Z}}\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{X}}\right\} = \Gamma^{1/2}E\left\{\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{X}}\right\}. \tag{A.3}$$

Hence the combination of (A.1)–(A.3) yields

$$\Gamma^{-1}\operatorname{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\} = E\left\{\frac{\partial m(\boldsymbol{X})}{\partial \boldsymbol{X}}\right\},$$

namely, $\boldsymbol{\theta} = E\{\partial m(\boldsymbol{X})/\partial \boldsymbol{X}\}$. Define $D_j = \partial m(\boldsymbol{X})/\partial X_j$, which gives

$$\theta_j = E(D_j). \tag{A.4}$$

Thus if $m(\boldsymbol{x})$ is monotone increasing with x_j, then $D_j \geq 0$ a.e. and thus $\theta_j \geq 0$. Moreover, since D_j is a non-negative random variable, it can be shown that $P(D_j > 0) > 0$ is equivalent to $E(D_j) > 0$. In addition, (A.4) implies the equivalence between the strict inequalities $E(D_j) > 0$ and $\theta_j > 0$. Similar arguments can be obtained when $m(\boldsymbol{x})$ is monotone decreasing with x_j. This completes the proof.

A.2. Proof of Theorem 2

In our proof, we shall need the following definition.

Definition 2. Two functions F and G are called "*similarly ordered*" if $\{F(x) - F(y)\}\{G(x) - G(y)\} \geq 0$ for all x in the domain of F and all y in the domain of G.

The notion of "*similarly ordered*" means that the corresponding functions are monotone in the same direction.

We now prove Theorem 2. Without loss of generality, we write $\boldsymbol{X} = (X_j, \boldsymbol{X}_{-j}^{\mathrm{T}})^{\mathrm{T}}$ and correspondingly $\boldsymbol{\theta} = (\theta_j, \boldsymbol{\theta}_{-j}^{\mathrm{T}})^{\mathrm{T}}$. If X_j is independent of $\boldsymbol{X}_{-j}$, then the matrix Γ in (2.1) reduces to the form,

$$\Gamma = \begin{bmatrix} \operatorname{var}(X_j) & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{0} & \operatorname{cov}(\boldsymbol{X}_{-j}, \boldsymbol{X}_{-j}) \end{bmatrix}, \tag{A.5}$$

where $\boldsymbol{0} = (0, \ldots, 0)^{\mathrm{T}}$, and by definition we can write

$$\theta_j = \{\operatorname{var}(X_j)\}^{-1}\operatorname{cov}\{X_j, m(\boldsymbol{X})\}. \tag{A.6}$$

Thus, the sign of θ_j equals the sign of $\operatorname{cov}\{X_j, m(\boldsymbol{X})\}$.

It follows that

$$\operatorname{cov}\{X_j, m(\boldsymbol{X})\} = E\{(X_j - \mu_j)m(\boldsymbol{X})\} = E(B_j). \tag{A.7}$$

Thus, we only need to discuss the sign of B_j in (A.7). Since X_j is independent of $\boldsymbol{X}_{-j}$, it follows from [3, p. 92] that the equality,

$$B_j = [E\{(X_j - \mu_j)m(X_j, \boldsymbol{x}_{-j})\}]|_{\boldsymbol{x}_{-j}=\boldsymbol{X}_{-j}}, \tag{A.8}$$

holds with probability 1. This is to say that $\boldsymbol{X}_{-j}$, which is a random sub-vector of B_j, can simply be treated as a deterministic sub-vector.

We only prove result (i), but result (ii) can be shown similarly. Assume that $m(\boldsymbol{x})$ is monotone increasing with x_j. On the right-side of (A.8), for any fixed $\boldsymbol{x}_{-j}$, it is easy to verify that $x_j - \mu_j$ and $m(x_j, \boldsymbol{x}_{-j})$, as univariate functions of x_j, are "*similarly ordered*". By the Tchebychef's inequality ([10], p. 43 and 168), we observe that for any fixed $\boldsymbol{x}_{-j}$,

$$E\{(X_j - \mu_j)m(X_j, \boldsymbol{x}_{-j})\} \geq E(X_j - \mu_j)\, E\{m(X_j, \boldsymbol{x}_{-j})\} = 0. \tag{A.9}$$

It follows from (A.8) and (A.9) that $B_j \geq 0$ a.e. Thus $E(B_j) \geq 0$. This combined with (A.6) and (A.7) leads to $\theta_j \geq 0$.

We now study conditions for the strict inequality, $\theta_j > 0$. For a non-negative random variable B_j, it can be shown that $P(B_j > 0) > 0$ is equivalent to $E(B_j) > 0$ which, according to (A.6) and (A.7), is equivalent to $\theta_j > 0$. The proof is completed. □

A.3. *Proof of Theorem 3*

Since $\partial m(\boldsymbol{x})/\partial \boldsymbol{x} = \boldsymbol{\beta}$ for a linear function $m(\boldsymbol{x})$, the monotone direction of $m(\boldsymbol{x})$ with respect to x_j is completely determined by the sign of β_j. Furthermore, it is straightforward to see that

$$\operatorname{cov}\{\boldsymbol{X}, m(\boldsymbol{X})\} = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{X}, \boldsymbol{X})\boldsymbol{\beta} = \Gamma\boldsymbol{\beta},$$

thus by definition,

$$\boldsymbol{\theta} = \Gamma^{-1}\Gamma\boldsymbol{\beta} = \boldsymbol{\beta}.$$

This completes the proof.

A.4. *Proof of Theorem 4*

Condition (C1) implies that the matrix Γ in (2.1) reduces to the form (A.5) and (A.6) continues to hold. Thus, the sign of θ_j equals the sign of $\operatorname{cov}\{X_j, m(\boldsymbol{X})\}$. It follows immediately that

$$\begin{aligned}\operatorname{cov}\{X_j, m(\boldsymbol{X})\} &= \operatorname{cov}\{X_j, m_j(X_j)\} + \operatorname{cov}(X_j, \boldsymbol{\beta}_{-j}^{\mathrm{T}}\boldsymbol{X}_{-j}) \\ &= \operatorname{cov}\{X_j, m_j(X_j)\},\end{aligned}$$

where the first equality is due to condition (C2) and the second equality is due to condition (C1). Hence, the sign of θ_j equals the sign of $\operatorname{cov}\{X_j, m_j(X_j)\}$.

We only show result (i), but result (ii) can be proved similarly. If $m(\boldsymbol{x})$ is monotone increasing with the jth coordinate x_j, then this is equivalent to say, from assumption (C2), that $m_j(x_j)$ is monotone increasing with x_j. In this univariate case, it follows from Remark 2 that $\operatorname{cov}\{X_j, m_j(X_j)\} \geq 0$; moreover, in this case, $\operatorname{cov}\{X_j, m_j(X_j)\} > 0$ if and only if $P\{(X_j - \mu_j)m_j(X_j) > 0\} > 0$. The proof is completed.

A.5. Proof of Theorem 5

Condition (D1) implies that X_j is un-correlated with $\boldsymbol{X}_{-j}$. Again, the matrix Γ in (2.1) reduces to the form (A.5) and (A.6) continues to hold. Thus, the sign of θ_j equals the sign of $\mathrm{cov}\{X_j, m(\boldsymbol{X})\}$. It follows immediately that

$$\begin{aligned}\mathrm{cov}\{X_j, m(\boldsymbol{X})\} &= \mathrm{cov}\{X_j, m_j(X_j)\} + \mathrm{cov}\{X_j, M_j(\boldsymbol{X}_{-j})\} \\ &= \mathrm{cov}\{X_j, m_j(X_j)\},\end{aligned}$$

where the first equality is due to condition (D2) and the second equality is due to condition (D1). Hence, the sign of θ_j equals the sign of $\mathrm{cov}\{X_j, m_j(X_j)\}$.

The rest of the proof resembles the arguments used in Theorem 4 and is omitted. The proof is completed.

References

[1] S.C. Albright, W.L. Winston, C.J. Zappe, Data Analysis and Decision Making with Microsoft Excel, Duxbury Press, Pacific Grove, CA, 1999.
[2] J.H. Cochrane, Asset Pricing, Princeton University Press, Princeton, 2001.
[3] A.Ya. Dorogovtsev, D.S. Silvestrov, A.V. Skorokhod, M.I. Yadrenko, Probability Theory: Collection of Problems, American Mathematical Society, Providence, RI, 1997.
[4] D. Drouet Mari, S. Kotz, Correlation and Dependence, Imperial College Press, London, 2001.
[5] R.L. Dykstra, T. Robertson, An algorithm for isotonic regression for two or more independent variables, Ann. Statist. 10 (1982) 708–716.
[6] K.T. Fang, S. Kotz, K.W. Ng, Symmetric Multivariate and Related Distributions, Chapman & Hall, London, 1990.
[7] K.A. Froot, 2003. Risk management, capital budgeting and capital structure policy for insurers and reinsurers, J. Risk Insurance 2007 (in press). See http://ideas.repec.org/p/nbr/nberwo/10184.html#download.
[8] F. Gebhardt, An algorithm for monotone regression with one or more independent variables, Biometrika 57 (1970) 263–271.
[9] W. Härdle, H. Liang, J.T. Gao, Partially Linear Models, Physica-Verlag, Heidelberg, 2000.
[10] G.H. Hardy, J.E. Littlewood, G. Pólya, Inequalities, second edition, Cambridge University Press, Cambridge, England, 1988.
[11] T.J. Hastie, R.J. Tibshirani, Generalized Additive Models, Chapman and Hall, London, 1990.
[12] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer-Verlag, New York, 2001.
[13] Z. Landsman, On the generalization of Stein's Lemma for elliptical class of distributions, Statist. Probab. Lett. 76 (2006) 1012–1016.
[14] Z. Landsman, J. Nešlehová, Stein's Lemma for elliptical random vectors, J. Multivariate Anal. 99 (5) (2008) 912–927.
[15] E. Langford, N. Schwertman, M. Owens, Is the property of being positively correlated transitive?, Amer. Statist. 55 (2001) 322–325.
[16] P. McCullagh, J.A. Nelder, Generalized Linear Models, 2nd ed., Chapman and Hall, London, 1989.
[17] H. Mukarjee, S. Stern, Feasible nonparametric estimation of multiargument monotone functions, J. Amer. Statist. Assoc. 89 (1994) 425. 77–80.
[18] J.A. Nelder, R.W.M. Wedderburn, Generalized linear models, J. Roy. Statist. Soc. Ser. A 135 (1972) 370–384.
[19] H.H. Panjer (Ed.), Financial Economics: With Applications to Investments, Insurance and Pensions, The Actuarial Foundation, Schaumberg, IL, 1998.
[20] W. Rudin, Principles of Mathematical Analysis, third edition, McGraw-Hill Book Co., New York-Auckland-Dsseldor, 1976.
[21] C.M. Stein, Estimation of the mean of a multivariate normal distribution, Ann. Statist. 9 (1981) 1135–1151.
[22] M.P. Wand, A central limit theorem for local polynomial backfitting estimators, J. Multivariate Anal. 70 (1999) 57–65.
[23] C.M. Zhang, Prediction error estimation under Bregman divergence for non-parametric regression and classification, Scand. J. Statist. (in press).
[24] Z.J. Zhang, Z.Q. Yang, C.M. Zhang, Monotone piecewise curve fitting algorithms, J. Comput. Math. 12 (1994) 163–172.