Statistica Sinica **25** (2015), 787-807
doi:http://dx.doi.org/10.5705/ss.2013.168

SUFFICIENT DIMENSION REDUCTION FOR LONGITUDINAL DATA

Xuan Bi and Annie Qu

University of Illinois at Urbana-Champaign

Abstract: Correlation structure contains important information about longitudinal data. Existing sufficient dimension reduction approaches assuming independence may lead to substantial loss of efficiency. We apply the quadratic inference function to incorporate the correlation information and apply the transformation method to recover the central subspace. The proposed estimators are shown to be consistent and more efficient than the ones assuming independence. In addition, the estimated central subspace is also efficient when the correlation information is taken into account. We compare the proposed method with other dimension reduction approaches through simulation studies, and apply this new approach to longitudinal data for an environmental health study.

Key words and phrases: Correlation structure, eigen-decomposition, quadratic inference function, slice inverse regression, transformation method.

1. Introduction

Sufficient dimension reduction plays an important role in reducing the dimension of predictors and providing better modeling for response variables. The essential idea is to construct low-dimensional variables which can predict the response without loss of information. In contrast to the variable selection strategy, sufficient dimension reduction does not select or eliminate variables in a certain way. Instead, it extracts important information through optimally combining all predictors. Another advantage of sufficient dimension reduction is that it can be an effective way to visualize data (Li (1991)) through plotting the responses against the first several optimal combinations of covariates, which is especially important for handling high-dimensional data. Moreover, sufficient dimension reduction provides essential tools in analysis and curation for high-dimensional data, as it is able to reduce the original high-dimension of data to a moderate size without losing important information.

Existing methods of sufficient dimension reduction include, but are not limited to, ordinary least square (OLS; Li and Duan (1989)), slice inverse regression (SIR; Li (1991)), sliced average variance estimation (SAVE; Cook and Weisberg (1991)), principal Hessian direction (PHD; Li (1992)), discriminant analysis

(Cook and Yin (2001); Pardoe, Yin, and Cook (2007)), minimum average variance estimation (MAVE; Xia et al. (2002)), coutour regression (CR; Li, Zha, and Chiaromonte (2005)), inverse regression estimation (IRE; Cook and Ni (2005)), directional regression (DR; Li and Wang (2007)), sliced regression (SR; Wang and Xia (2008)), contour projection (CP; Luo, Wang, and Tsai (2009)), dimension reduction for non-elliptically distributed predictors (Li and Dong (2009); Dong and Li (2010)), and dimension reduction based on canonical correlation (Fung et al. (2002); Zhou and He (2008); Zhou (2009)). The study of sufficient dimension reduction for longitudinal data is still quite limited. With the prevalence of longitudinal study in biomedical, social, political, psychological, and environmental sciences, and with the increasing demand for handling high-dimensional data, it is of great importance to address sufficient dimension reduction problems under the longitudinal data framework.

For the longitudinal data setting, following Li, Cook, and Chiaromonte's partial OLS (2003) Li and Yin (2009) propose an analog partial OLS by conducting OLS at each time point and extracting a small subset of eigenvectors to achieve longitudinal data dimension reduction. However, their method does not incorporate intracluster correlation structure, and therefore leads to a significant loss of correlation information. In addition, their method is not able to exhaust the central subspace (Cook and Weisberg (1994); Cook (1996, 1998)) if the cluster size is less than the structural dimension. Pfeiffer, Forzani, and Bura (2012) propose a longitudinal first-moment-based sufficient dimension reduction method to solve these problems. They utilize a Kronecker-product space of clusters and predictors, and successfully accommodate the correlation structure of longitudinal covariates. However, their method is mainly applicable for handling longitudinal covariates, and not for longitudinal responses.

In this paper, we apply the quadratic inference function (QIF; Qu, Lindsay, and Li (2000)) to longitudinal data sufficient dimension reduction, which can accommodate both longitudinal responses and correlation information. The QIF improves the generalized estimating equation (GEE; Liang and Zeger (1986)) without estimating nuisance parameters, and is shown to be efficient in regression parameter estimation for longitudinal data. In our approach, we first identify a group of transformation functions for the responses, then minimize the quadratic inference function which incorporates correlation information for transformed responses to obtain regression parameter estimators, and then apply eigen-decomposition to extract information from a set of regression parameter estimators for the transformed responses.

The proposed method allows one to gain extra efficiency in parameter estimation for both continuous and discrete responses through incorporating correlation structure, while not requiring that the true correlation structure be known. Most importantly, we obtain parameter estimation from the entire cluster instead of performing regression separately at each time point, as in Li and Yin (2009). This leads to several advantages, such that the proposed method can still be efficient even for a small sample size, since we utilize information from repeated measurements within the same subject, and therefore the sample points used in our estimation are larger than the ones in Li and Yin (2009). In addition, the proposed method is computationally more efficient than existing methods, as the operation cost is lower for the same reason. In our approach the recovery of the central subspace does not depend on the cluster size, is in contrast to existing approaches which require the cluster size to be greater than the structural dimension.

In theory, we show that estimation through minimizing the QIF for the transformed data is still in the central subspace, and asymptotic efficiency can be improved by incorporating correlation structures. Another finding is that the efficiency of parameter estimation leads to the efficiency of the central subspace estimation. This is confirmed by our simulation studies, which show that the proposed method can improve accuracy and efficiency for sufficient dimension reduction in finite samples.

The remainder of the paper is organized as follows. Section 2 provides background for the quadratic inference function. Section 3 introduces the proposed method for longitudinal dimension reduction using the QIF, and provides its theoretical foundation and properties. Section 4 illustrates how to recover the structural dimension and provides the implementation of the proposed method. Section 5 compares the proposed approach with existing work through simulation studies for normal and binary responses. Section 6 applies the proposed method to a longitudinal asthma study. The last section concludes our findings and provides a brief discussion. Technical derivations are provided in the Appendix.

2. Quadratic Inference Function

For longitudinal data, suppose y_{it} is the response of subject i at time t, and $\mathbf{x}_{it}$ is a p-dimensional covariate, where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. To simplify notation, we set $T_i = T$ for all i; the unbalanced data case will be discussed in more details in Section 4. Let $\mu(\cdot)$ be an inverse link function satisfying $\mathrm{E}(y_{it}|\mathbf{x}_{it}) = \mu(\boldsymbol{\beta}'\mathbf{x}_{it})$, where $\boldsymbol{\beta}$ is a p-dimensional parameter. Define $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{y}_i|\mathbf{x}_i)$ for each i. If independence structure is assumed among subjects, the quasi-likelihood equation (Wedderburn (1974); McCullagh (1983)) for solving $\boldsymbol{\beta}$ is

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\dot{\boldsymbol{\mu}}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is a $T \times p$ matrix, and $\mathbf{V}_i$ is the covariance matrix of $\mathbf{y}_i$. In practice $\mathbf{V}_i$ is usually unknown. One common approach is to substitute the empirical estimator $\hat{\mathbf{V}}_i$ for $\mathbf{V}_i$. However, this involves many nuisance parameter estimations and thus $\hat{\mathbf{V}}_i$ can be unstable when T is large. Liang and Zeger (1986) introduced the working correlation matrix which reduces the number of correlation parameters significantly. They assume $\widetilde{\mathbf{V}}_i = \mathbf{A}_i^{1/2}\mathbf{R}(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is a diagnal matrix of marginal variance of $\mathbf{y}_i$, $\mathbf{R}(\boldsymbol{\alpha})$ is the working correlation matrix, and $\boldsymbol{\alpha}$ contains a small number of correlation parameters.

The QIF approach (Qu, Lindsay, and Li (2000)) further avoids the estimation of $\boldsymbol{\alpha}$ by formulating $\mathbf{R}^{-1}$ as a linear combination of $\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_{m-1}$, where $\mathbf{M}_0$ is a T-dimensional identity matrix. For example, if $\mathbf{R}(\boldsymbol{\alpha})$ is exchangeable, then $m = 2$ and $\mathbf{M}_1$ has 0 on the diagonal and 1 elsewhere. The idea of the QIF is to ensure the additional moment conditions $\sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-1/2}\mathbf{M}_r\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ are as close to 0 as possible for $r = 1, \ldots, m-1$. Since the number of equations is greater than the number of parameters, the QIF utilizes the generalized method of moments (GMM; Hansen (1982)), where the specified moment conditions of $\mathbf{b} \in \mathbb{R}^p$ for estimating $\boldsymbol{\beta}$ are

$$\mathbf{g}_i(\mathbf{b}) = \begin{pmatrix} (\dot{\boldsymbol{\mu}}_i)' \mathbf{A}_i^{-\frac{1}{2}}\mathbf{M}_0\mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ (\dot{\boldsymbol{\mu}}_i)' \mathbf{A}_i^{-\frac{1}{2}}\mathbf{M}_{m-1}\mathbf{A}_i^{-\frac{1}{2}}(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}, i = 1 \ldots, n. \tag{2.1}$$

The quadratic inference function is defined as

$$\hat{Q}(\mathbf{b}) = n\bar{\mathbf{g}}'(\mathbf{b})\hat{\mathbf{W}}^{-1}(\mathbf{b})\bar{\mathbf{g}}(\mathbf{b}), \tag{2.2}$$

where $\bar{\mathbf{g}}(\mathbf{b}) = 1/n\sum_{i=1}^n \mathbf{g}_i(\mathbf{b})$, and $\hat{\mathbf{W}}(\mathbf{b}) = 1/n\sum_{i=1}^n \mathbf{g}_i(\mathbf{b})\mathbf{g}_i'(\mathbf{b})$. The corresponding QIF estimator is obtained as $\hat{\mathbf{b}} = \text{argmin}_b \hat{Q}(\mathbf{b})$. Qu, Lindsay, and Li (2000) showed that $\hat{\mathbf{b}}$ is a $\sqrt{n}$-consistent estimator and is efficient if a linear combination of basis matrices $\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_{m-1}$ contains the true correlation structure.

A critical issue regarding the QIF is the selection of the number m of basis matrices, which has been addressed by model selection for correlation structure in Zhou and Qu (2012). The basic idea is to approximate the inverse of the empirical correlation matrix by a group of basis matrices, which contain only 0 and 1 as entries. Then a Euclidean-norm measuring the difference between two estimating functions, one based on the empirical correlation information and the other on the model-based approximation, is minimized. Through a groupwise penalty on the basis matrices, an appropriate number m of basis matrices can be selected such that sufficient correlation information is captured. In theory, the selected correlation structure is consistent if the candidate basis matrices are from a sufficiently rich class to represent the true structure.

In general, the moment condition $\mathbf{g}_i(\mathbf{b}) = \mathbf{g}(\mathbf{b}'\mathbf{x}_i, \mathbf{y}_i)$ is required to satisfy $\mathrm{E}(\mathbf{g}_i) = \mathbf{0}, i = 1, \ldots, n$, to identify the true parameter $\boldsymbol{\beta}$. The population version of the QIF is $Q(\mathbf{b}) = (\mathrm{E}\mathbf{g})'\mathbf{W}^{-1}(\mathrm{E}\mathbf{g})$, where $\mathbf{W} = \mathrm{Var}(\mathbf{g})$. Therefore, $Q(\mathbf{b}) \geq 0$, and the equality holds if and only if $\mathbf{b} = \boldsymbol{\beta}$.

3. Sufficient Dimension Reduction for Longitudinal Data

In this section, we propose the QIF approach for sufficient dimension reduction in the longitudinal data setting.

Let $\mathbf{X}$ be a $p \times T$-dimensional covariate matrix and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)'$ be a T-dimensional response. Both $\mathbf{X}$ and $\mathbf{Y}$ can be random. The main purpose of sufficient dimension reduction (SDR; Li (1991); Cook (1998)) is to seek a minimal dimension-reduction subspace with a $p \times d$ basis matrix $\mathbf{B}$, where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$, $d \leq p$, such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$. Here $\perp\!\!\!\perp$ indicates independence. Under some regularity conditions (Cook (1998)), the minimal subspace exists and is unique, that is, the central subspace of the regression of $\mathbf{Y}$ on $\mathbf{X}$, denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Suppose rank($\mathbf{B}$)= d, then $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is also called the structural dimension of regression. The central subspace is the smallest subspace of $\mathbb{R}^p$ that captures all of the regression information of $\mathbf{Y}$ given $\mathbf{X}$, and therefore reduces the dimension of the predictors from $\mathbf{X}$ to $\mathbf{B}'\mathbf{X}$.

We propose to identify the central subspace by recovering its basis through minimizing the QIF. If the dimension of the central subspace is $d = 1$, then the problem of identifying the central subspace is equivalent to a parameter estimation problem, and thus the QIF estimator alone can capture the central subspace completely, due to the fact that $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{Span}(\boldsymbol{\beta}_1)$. When $d \geq 2$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ may not be identifiable (Li (1991)).

Alternatively, to recover the central subspace, we propose to minimize the QIF for transformed responses. This approach does not have the identifiability constraint for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$. Suppose we have a group of transformations h_j's for responses, $h_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, s$. Let $\mathbf{h}_j = (h_j, \ldots, h_j)'$ be a T-dimensional transformation function vector on the response vector $\mathbf{Y}$. Take

$$Q_j(\mathbf{b}) = \left\{\mathrm{E}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{h}_j(\mathbf{Y}))\right\}' \mathbf{W}^{-1} \left\{\mathrm{E}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{h}_j(\mathbf{Y}))\right\}, \tag{3.1}$$

with minimizer $\boldsymbol{\gamma}_j = \mathrm{argmin}_b\, Q_j(\mathbf{b})$, $j = 1, \ldots, s$. In Section 3.1, we show that $\boldsymbol{\gamma}_j$ is in the central subspace under certain conditions, and $\mathrm{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)$ approximates $\mathcal{S}_{Y|\mathbf{X}}$. Since d is typically unknown, we need a sufficiently large s to ensure that $s \geq d$. The selection of s is discussed in Section 4.2 in detail.

There are several strategies to choose the transformation function h_j. One common practice is to use the power transformation (Cook and Li (2002);

Yin and Cook (2002); Zhu and Zhu (2009); Yin and Li (2011)), $h_j(Y_t) = Y_t^j, j = 1, \ldots, s$. Other transformation methods include the slice indicator function proposed by Li (1991), which defines $h_j(Y_t) = 1$ if Y_t is in the jth slice and 0 otherwise, the covariance inverse regression method (Cook and Ni (2006)) defining $h_j(Y_t) = Y_t$ if Y_t is in the jth slice and 0 otherwise, and the normalized B-spline basis functions for Y_t (Fung et al. (2002)). Cook (1998, p.114) shows that $\mathcal{S}_{h(Y)|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ holds for any transformation function h, and $\mathcal{S}_{h(Y)|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$ holds if h is a one-to-one function.

The purpose of applying the transformation method is that, although minimizing the QIF from the original responses can only recover one basis vector for the central subspace, the transformation method can provide a group of transformed responses, and therefore recover a group of basis vectors that allow one to explore the central subspace to its largest extent.

3.1. Theoretical properties

We assume the well-known *linearity condition* (Li and Duan (1989)) that states that $\mathrm{E}(\mathbf{X}|\mathbf{B}'\mathbf{X})$ is linear in $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_d'\mathbf{X}$. This entails that the distribution of $\mathbf{X}$ be elliptically symmetric. Li and Dong (2009), Dong and Li (2010) and Ma and Zhu (2012, 2013a,b) provide alternative strategies on how to relax this condition. On the other hand, the *constant conditional variance* assumption (Cook and Weisberg (1991)), where $\mathrm{Var}(\mathbf{X}|\mathbf{B}'\mathbf{X})$ is a constant matrix, is not required.

Suppose $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$, and let $L(\mathbf{b}'\mathbf{X}, \mathbf{Y}) = \mathbf{g}'(\mathbf{b}'\mathbf{X}, \mathbf{Y})\mathbf{W}^{-1}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{Y})$ be a loss function. Then the following theorem shows that the QIF minimizer,

$$\boldsymbol{\gamma} = \operatorname*{argmin}_{b} Q(\mathbf{b}), \boldsymbol{\gamma} \in \mathbb{R}^p, \tag{3.2}$$

is in the central subspace. In addition, the sample estimator

$$\hat{\boldsymbol{\gamma}} = \operatorname*{argmin}_{b} \hat{Q}(\mathbf{b}) \tag{3.3}$$

is a strongly consistent estimator of $\boldsymbol{\gamma}$, where $\hat{Q}(\mathbf{b})$ is defined in (2.2).

Theorem 1. *Assume $L(\cdot, \cdot)$ is convex in its first argument, the linearity condition holds, and Var$(\mathbf{X})$ is positive definite. If $\boldsymbol{\gamma}$ in* (3.2) *exists and is unique, then $\boldsymbol{\gamma} \in \mathcal{S}_{Y|\mathbf{X}}$, and $\hat{\boldsymbol{\gamma}}$ in* (3.3) *converges to $\boldsymbol{\gamma}$ almost surely.*

The convexity condition of $L(\cdot, \cdot)$ in its first argument is easily satisfied in our approach since $\ddot{L} = \dot{\mathbf{g}}'\mathbf{W}^{-1}\dot{\mathbf{g}} + o_p(1)$ is a non-negative definite matrix, asymptotically. The strict convexity of L is a sufficient condition to ensure the uniqueness of $\boldsymbol{\gamma}$ in Theorem 1 (Li and Duan (1989)).

When $d = 1$ and $\mathbf{g}_i$ is defined in equation (2.1), Theorem 1 implies that the minimizer $\boldsymbol{\gamma}$ in (3.2) is the true parameter, and the sample minimizer $\hat{\boldsymbol{\gamma}}$ in (3.3) converges to the true parameter $\boldsymbol{\gamma}$ almost surely.

Theorem 1 does not require $\mathbf{g}$ to satisfy $\mathrm{E}(\mathbf{g}) = 0$, the strong consistency property is robust to the misspecification of the link functions. This is even more desirable when the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is difficult to find. As for the efficiency argument in Section 3.2, however, a correctly specified link function is required to achieve an efficiency gain through incorporating correlation information.

Theorem 1 lays the foundation for formulating basis vectors for the central subspace. Suppose $\hat{Q}_j(\mathbf{b})$ is the sample version of $Q_j(\mathbf{b})$, and $\hat{\boldsymbol{\gamma}}_j = \mathrm{argmin}_b\, \hat{Q}_j(\mathbf{b})$ is the sample estimator of $\boldsymbol{\gamma}_j$.

Corollary 1. *Assume $L(\cdot,\cdot)$ is convex in its first argument, the linearity condition holds, and Var($\mathbf{X}$) is positive definite. If $\boldsymbol{\gamma}_j$ exists and is unique, then $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$, and $\hat{\boldsymbol{\gamma}}_j$ converges to $\boldsymbol{\gamma}_j$ almost surely, $j = 1, \ldots, s$.*

Corollary 1 implies that each $\boldsymbol{\gamma}_j$ is a linear combination of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$, and that $\mathrm{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. This provides an effective way to build a central subspace basis. If $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d$ are the eigenvectors corresponding to the largest d eigenvalues of $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)\ (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)'$, then the basis for the central subspace can be taken as $\hat{\mathbf{B}} = (\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d)$.

Recently, Yin and Li (2011) formulated the conditions to achieve exhaustiveness of the central subspace, which can accommodate power transformations as a special case. In their Theorem 2.1 and Example 2.1, they proved that given a sufficiently large s, the subspace spanned by $\mathcal{S}_{E(Y^j|\mathbf{X})}$ $(j = 1, \ldots, s)$ approaches the central subspace under mild conditions, where $\mathcal{S}_{E(Y|\mathbf{X})}$ denotes the central mean subspace of $\mathbf{Y}$ on $\mathbf{X}$ (Cook and Li (2002)). For each transformation $\mathbf{Y}^j$, the quadratic inference function (QIF) can recover one basis vector from $\mathcal{S}_{E(Y^j|\mathbf{X})}$. Therefore, a sufficient condition to achieve exhaustiveness, as mentioned in Yin and Cook (2002), is to assume that there exists a group of powers $k_1, \ldots, k_d$, such that $\dim(\mathcal{S}_{E(Y^{k_j}|\mathbf{X})}) = 1$ for $j = 1, \ldots, d$. Under such an assumption, the QIF approach with the transformed response $\mathbf{Y}^j, j = 1, 2, \ldots, k_d$, can exhaust the central subspace. When other types of tranformations are applied, a similar assumption should be satisfied accordingly. Exhaustiveness can then be achieved if the new tranformations follow the conditions of Theorem 2.1 in Yin and Li (2011). Alternatively, Ma and Zhu (2012, 2013c) propose a semiparametric estimating equation approach that avoids the aforementioned condition, but still achieves exhaustiveness by identifying and estimating the central subspace basis $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)$ using one estimating equation.

3.2. Efficiency

As long as the responses from the same subject are not independent, incorporating correlation information always leads to efficiency gain. In addition, the efficiency gain of parameter estimation from the data with each transformation of the response variable provides an overall efficiency gain of the central subspace estimation.

For illustration, suppose there are two sets of moment conditions: $\mathbf{G}_l = \sum_{i=1}^{n}(\dot{\boldsymbol{\mu}}_i)'\mathbf{A}_i^{-1/2}\ \mathbf{M}_l\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i)$, $l = 1, 2$, where $\mathbf{M}_1$ and $\mathbf{M}_2$ are symmetric matrices, $\dot{\boldsymbol{\mu}}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is the true parameter. Let $\mathbf{G} = (\mathbf{G}_1', \mathbf{G}_2')'$, $\dot{\mathbf{G}} = \partial\mathbf{G}/\partial\boldsymbol{\beta}$, $\dot{\mathbf{G}}_1 = \partial\mathbf{G}_1/\partial\boldsymbol{\beta}$, $\mathbf{C} = \mathrm{Var}(\mathbf{G})$, and $\mathbf{C}_{11} = \mathrm{Var}(\mathbf{G}_1)$. The empirical information matrices corresponding to $\mathbf{G}$ and $\mathbf{G}_1$ are $\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}}$ and $\dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$, respectively. We show that incorporating a correlation structure leads to an increase of the empirical information matrix in the sense of the Loewner ordering (Beckenback and Bellman (1965)), which is equivalent to an improvement in parameter estimation efficiency.

Lemma 1. *If* $\mathbf{R}^{-1} = a_1\mathbf{M}_1 + a_2\mathbf{M}_{\mathbf{2}}$ *is the true correlation matrix and* $E(\mathbf{G}) = \mathbf{0}$*, then* $\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} \geq \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$*, in terms of the Loewner ordering for matrices. Equality holds if* $a_2 = 0$*.*

Lemma 1 indicates that we gain efficiency by incorporating additional correlation information; if $\mathbf{M}_1$ is an identity matrix, then the proposed dimension reduction method incorporating correlation structure is more efficient than those assuming independence. In simulation studies provided in Section 5, we illustrate that the performance of sufficient dimension reduction based on the QIF assuming independence is similar to other approaches such as the OLS or SIR, while the QIF incorporating correlation information can significantly improve the efficiency for sufficient dimension reduction.

The condition $\mathrm{E}(\mathbf{G}) = \mathbf{0}$ assumes that each moment condition has zero expectation. That is, use the conditional mean $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ as a link function for the transformed response $\mathbf{h}_j(\mathbf{Y})$ (Yin and Cook (2002); Ma and Zhu (2012, 2013b)). In practice, however, the conditional mean $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ is usually unknown. As pointed out by Ma and Zhu (2013c), unless one uses a nonparametric approach, it might be difficult to find the correct link function. For the proposed method, this is even more challenging than in Ma and Zhu's (2013c) case, since for each transformation the QIF can only generate one basis vector for the central subspace. Unless we assume $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ is known and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ are identifiable, the link function of the QIF is typically misspecified.

A possible way to have a correctly specified link function might be to apply a nonparametric procedure, but this could complicate our method significantly. To avoid this, Ma and Zhu (2013c) also suggested imposing additional assumptions;

for example, the linearity condition on $\mathbf{Y}$, or applying a common link function. In one simulation study, we apply a common (identity) link function. There are two practical justifications for this application. First, the response $\mathbf{Y}$ is continuous and could range from negative infinity to infinity. Second, using the identity link is a linear approximation of the true link function. Thus, even though the link function may not be exact, it will still achieve good efficiency in practice. In fact, we find that the proposed method with the identity link function indeed has an efficiency gain through incorporating correlation information. Other common link functions can be applied when the response is not continuous. Refer to Ma and Zhu (2013c,d) for more detail.

The consistency of the estimator for a central subspace vector is guaranteed by Corollary 1, and the efficiency gained by incorporating correlation information can be followed by Lemma 1.

Theorem 2. *Suppose $\hat{\boldsymbol{\gamma}}_j$ is an efficient estimator of $\boldsymbol{\gamma}_j$ corresponding to the j-th transformation function, where $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$, $j = 1, \ldots, s$. Then $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)$ is an efficient estimator of $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)$, provided the information matrix corresponding to the true parameter $(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$ is bounded.*

4. Implementation

4.1. Estimation of structural dimension

For selection of structural dimension d, several approaches have been proposed. Li (1991) provided an asymptotic chi-squared test, assuming that the covariates are normally distributed, and Cook and Yin (2001) built the foundation of the permutation test for the structural dimension. In addition, Li and Wang (2007) introduced a sequential test, and Ye and Weiss (2003) proposed a bootstrap procedure. Luo, Wang, and Tsai (2009) further suggested a quick and effective selection procedure called the *maximal eigenvalue ratio criterion*, which chooses

$$\hat{d} = \underset{1 \le q \le d_{max}}{\text{argmax}} \frac{\hat{\lambda}_q}{\hat{\lambda}_{q+1}}. \tag{4.1}$$

In practice, $d_{max} = 5$ usually suffices. The intuition behind (4.1) can be explained. Suppose $\hat{\mathbf{B}}$ is a consistent estimator of $\mathbf{B}$, and therefore that each $\hat{\lambda}_q$ converges to λ_q consistently. Since $\dim(\mathbf{B}) = d$, λ_q's are nonzero if $q \le d$. As $\lim_{n\to\infty} \hat{\lambda}_d / \hat{\lambda}_{d+1} = +\infty$, choosing $\hat{d}$ to satisfy (4.1) is a sensible approach.

4.2. Algorithm

We provide an algorithm for sufficient dimension reduction for longitudinal data.

(i) Choose a transformation function $\mathbf{h}_j$, and transform the response $\mathbf{y}_i$ into $\mathbf{h}_j(\mathbf{y}_i)$, for $j = 1, \ldots, s$ and $i = 1, \ldots, n$.

(ii) For the transformed responses $\mathbf{h}_j(\mathbf{y}_1), \ldots, \mathbf{h}_j(\mathbf{y}_n)$, obtain $\hat{\boldsymbol{\gamma}}_j$ by minimizing $\hat{Q}_j(\mathbf{b})$.

(iii) Conduct a spectral decomposition for $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)'$, and obtain the structural dimension d based on (4.1).

(iv) Select eigenvectors $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d$ corresponding to the first to d-th largest eigenvalues, and formulate the basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

That the selection of s in (i) is similar to, but less critical than, the selection of the number of slices in SIR, is still an open question (Wang and Xia (2008)). If $\lim_{s\to\infty} \mathrm{Span}(\gamma_1, \ldots, \gamma_s) = \mathcal{S}_{Y|X}$, the transformed QIF with a sufficiently large s could approximate the central subspace (compared to SIR where the number of slices may be restricted if the support of $\mathbf{Y}$ is finite). On the other hand, a finite and fixed s may not be enough to exhaust the central subspace even if we are given $\lim_{s\to\infty} \mathrm{Span}(\gamma_1, \ldots, \gamma_s) = \mathcal{S}_{Y|X}$ (Yin and Li (2011)), as difficult as the SIR in choosing the total number of slices.

In practice, the selection of s may not be very critical, similar to the selection of the total number of slices for many inverse regression methods, e.g., SIR, SAVE and SR. Our numerical studies indicate that the proposed method is rather robust against s: the simulation results did not change much once $s \geq d$. Currently if d can be detected by other methods, as in our data analysis for the asthma study in Section 6, then s can be selected accordingly.

4.3. Implementation with unbalanced data

In practice, unbalanced data are quite common. If the measurements from unbalanced data are regarded as cluster data without considering the order of lag time, then each $\boldsymbol{\mu}_i$ is a T_i-dimensional vector, and $\mathbf{M}_r$ is a $T_i \times T_i$ matrix for $i = 1, \ldots, n$ and $r = 0, 1, \ldots, m-1$.

If the lag time between measurements is considered important, we can define $T = \max(T_1, \ldots, T_n)$, and impose a $T \times T_i$ dimensional transformation matrix $\mathbf{U}_i$ for the i-th subject. Let $\mathbf{y}_i^* = \mathbf{U}_i\mathbf{y}_i$, $\boldsymbol{\mu}_i^* = \mathbf{U}_i\boldsymbol{\mu}_i$, $\dot{\boldsymbol{\mu}}_i^* = \mathbf{U}_i\dot{\boldsymbol{\mu}}_i$ and $\mathbf{A}_i^* = \mathbf{U}_i\mathbf{A}_i\mathbf{U}_i'$. Thus, we transform the unbalanced data to artificial balanced data where each component of $\mathbf{U}_i$ is an indicator of whether the data is observed or missing. Then we formulate moment conditions as in (2.1) for the newly created balanced data. The QIF estimator from minimizing (2.2) still has the right properties if the data are missing completely at random. See Zhou and Qu (2012) for more details.

5. Simulation

We report on simulation studies to illustrate the performance of the proposed method and existing approaches for longitudinal data sufficient dimension reduction. They show that incorporating a suitable correlation structure can improve the accuracy and efficiency of estimation for both the parameters and the central subspace.

5.1. Study 1: Binary responses with one set of parameters

We generated the covariate $\mathbf{x}_i$ as standard normal for subject $i = 1, \ldots, n$. For each $\mathbf{x}_i$, we assumed T repeated measurements $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and that each $\mathbf{x}_{it}$ is a p-dimensional vector, $t = 1, \ldots, T$. We assumed independence among different subjects and different covariates, but an exchangeable correlation structure among T time points for each covariate, with $\rho_x = 0.2$.

In Study 1, we let $p = 50, T = 20$, with sample size n as 51, 100, or 200. The true parameter $\boldsymbol{\beta}$ was a p-dimensional vector with 1 in its first 10 components and 0 otherwise. We generated $\mathbf{v}_i$ based on the linear model $\mathbf{v}_i = 0.4\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} N(0, \boldsymbol{\Sigma}_\varepsilon)$, $i = 1, \ldots, n$, where $\boldsymbol{\Sigma}_\varepsilon$ is a T-dimensional exchangeable correlation matrix with $\rho_\varepsilon = 0.2, 0.5$, or 0.8. We then generated $\mathbf{y}_i$ by utilizing an indicator function $y_{it} = \mathbf{1}_{A_{it}}$, where event $A_{it} = \{e^{v_{it}}/(1 + e^{v_{it}}) > 0.5\}$ and v_{it} is the t-th component of $\mathbf{v}_i$, $t = 1, \ldots, T$. Since $\mathbf{y}_i|\mathbf{x}_i = \mathbf{y}_i|\boldsymbol{\beta}'\mathbf{x}_i$, the structural dimension is $d = 1$, and the central subspace is $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}(\boldsymbol{\beta})$. It is straightforward that $\text{E}(y_{it}) = 0.5$ and $\text{E}(y_{it}|\mathbf{x}_{it}) = 1 - \Phi(0.4\boldsymbol{\beta}'\mathbf{x}_{it})$, where $\Phi(\cdot)$ is the standard normal distribution function. The correlation structure of $\mathbf{y}_i$ is close to, but not exactly, that of the exchangeable structure, as the correlation is mainly contributed by the error term $\boldsymbol{\varepsilon}_i$.

We measured the distance between central subspace basis matrix $\mathbf{B}$ and the estimated central subspace $\hat{\mathbf{B}}$ by $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$, where $||\cdot||_F$ is the Frobenius norm. We compared our method with the partial ordinary least square (partial OLS) by Li and Yin (2009), where the linear regression is conducted at each time point to recover parameter vectors for the central subspace, and d eigenvectors corresponding to the largest d eigenvalues are extracted through an eigen decomposition. We also compared with the "partial SIR," similar to Li and Yin's partial OLS except that at each time point linear regression is replaced by sliced inverse regression. Our simulation study shows that the partial SIR provides results similar to those of the partial OLS approach.

We generated simulation samples N =1,000. Table 1 provides the average distance, and the standard deviation (inside the parenthesis). The proposed dimension reduction method based on the QIF is significantly better than those from the partial OLS and partial SIR in the sense of accuracy and efficiency. For one, when $n = 51$ and $p = 50$, the partial OLS and the partial SIR provide estimators that are nearly orthogonal to the true parameter vector, while the proposed QIF is still robust, with much smaller distances between the true and estimated vectors. The linear regression at each time point has a sample size of 51, with a 50-dimensional parameter, so estimation is unstable. The proposed

Table 1. Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal binary data with $p = 50$ from 1,000 simulations.

		$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
			$n = 51$	
partial OLS	Independent	1.5507(0.1798)	1.5412(0.1632)	1.5603(0.1775)
partial SIR	Independent	1.5235(0.1955)	1.5209(0.1969)	1.5299(0.2007)
QIF	Independent	0.4142(0.0422)	0.4627(0.0497)	0.5070(0.0552)
	AR-1	0.4092(0.0416)	0.4260(0.0446)	0.4311(0.0436)
	Exchangeable	0.3981(0.0406)	0.3950(0.0416)	0.3871(0.0408)
			$n = 100$	
partial OLS	Independent	0.4208(0.0427)	0.4592(0.0468)	0.4963(0.0522)
partial SIR	Independent	0.4138(0.0423)	0.4521(0.0463)	0.4901(0.0517)
QIF	Independent	0.3008(0.0313)	0.3440(0.0371)	0.3833(0.0415)
	AR-1	0.2929(0.0300)	0.3073(0.0315)	0.3101(0.0305)
	Exchangeable	0.2830(0.0289)	0.2821(0.0287)	0.2752(0.0278)
			$n = 200$	
partial OLS	Independent	0.2427(0.0243)	0.2741(0.0270)	0.3029(0.0294)
partial SIR	Independent	0.2416(0.0243)	0.2731(0.0269)	0.3020(0.0292)
QIF	Independent	0.2153(0.0220)	0.2490(0.0251)	0.2792(0.0275)
	AR-1	0.2086(0.0212)	0.2193(0.0222)	0.2205(0.0216)
	Exchangeable	0.2005(0.0204)	0.1993(0.0203)	0.1931(0.0197)

method utilizes data from all time points simultaneously, so the number of sample points is $51 \times 20 = 1,020$, and this leads to a more precise estimation.

When the sample size is $n = 100$ or 200, the QIF assuming exchangeable correlation is still the best, though all methods converge to the true parameter space as the sample size increases. In an unreported simulation study, we found that the QIF converges faster than the other methods as the cluster size increases. The existing methods regress at each time point and have computing time dependent on the cluster size, while the QIF incorporates data from all time points simultaneously. As the cluster sizes increase, computational times of the proposed method and the existing approaches grow further apart.

Information on correlation has a strong influence on the estimations, and incorporating a correct correlation structure achieves higher accuracy and efficiency. The partial OLS and SIR approaches do not take correlation into account, and their results are relatively close to, but still worse than those estimated by the QIF dimension reduction approach assuming independence of data.

5.2. Study 2: Continuous responses with multiple sets of parameters

We investigated the performance of the new method when the dimension of central subspace d is greater than 1. Here we had two p-dimensional coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}_1'\mathbf{X}, \boldsymbol{\beta}_2'\mathbf{X})$, so $d = 2$. We

Table 2. Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model I from 1,000 simulations.

	Model I	$\rho_x = 0.2$	$\rho_x = 0.5$	$\rho_x = 0.8$
			$n = 100$, $2p = 16$	
partial OLS	Independent	1.1180(0.0730)	0.9337(0.1075)	0.8814(0.2544)
partial SIR	Independent	1.0652(0.1216)	1.2080(0.1394)	1.1708(0.1884)
QIF	Independent	0.4836(0.0265)	0.9060(0.0184)	1.1154(0.0105)
	AR-1	0.8142(0.0439)	0.6762(0.0454)	0.5903(0.0674)
	Exchangeable	0.8231(0.0322)	0.6409(0.0309)	0.5501(0.0365)
			$n = 300$, $2p = 30$	
partial OLS	Independent	1.0846(0.0378)	1.0163(0.0405)	1.1000(0.0830)
partial SIR	Independent	1.2093(0.0696)	1.2090(0.0996)	1.3272(0.0938)
QIF	Independent	0.9548(0.0304)	0.9450(0.0204)	1.0431(0.0133)
	AR-1	0.6930(0.0428)	0.7508(0.0447)	0.8741(0.0527)
	Exchangeable	0.5825(0.0358)	0.5883(0.0311)	0.6203(0.0368)

set $p = 8$ or 15. When $p = 8$, we let $\boldsymbol{\beta}_1 = (1,1,1,1,1,1,1,1)'/\sqrt{8}$, and $\boldsymbol{\beta}_2 = (1,-1,1,-1,1,-1,1,-1)'/\sqrt{8}$; when $p = 15$, we set the rest of the 7 components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to be 0. The continuous response variable y_{it} was generated using

$$\text{Model I: } y_{it} = \exp(\boldsymbol{\beta}_1'\mathbf{x}_{it}) + 2(1 + \boldsymbol{\beta}_2'\mathbf{x}_{it})^2 + 0.5(\boldsymbol{\beta}_1'\mathbf{x}_{it})\tau_{it};$$

$$\text{Model II: } y_{it} = \frac{(0.45\boldsymbol{\beta}_1'\mathbf{x}_{it})}{\{0.5 + (1.5 + \boldsymbol{\beta}_2'\mathbf{x}_{it})^2\}} + 0.5\varepsilon_{it};$$

$$\text{Model III: } y_{it} = \sin\left(\frac{\boldsymbol{\beta}_1'\mathbf{x}_{it}}{4}\right) + \exp\left(\frac{2\boldsymbol{\beta}_2'\mathbf{x}_{it}}{3}\right) + 0.5\varepsilon_{it}.$$

In Model I, we took ρ_x, the correlation of $\mathbf{x}_i$, as 0.2, 0.5, or 0.8, and took the error $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iT})' \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}_T)$, $i = 1, \ldots, n$. Because of heteroscedasticity, the responses in Model I are highly correlated, even though τ_{it}'s are independent. In Models II and III, we generated each $\mathbf{x}_i$ the same way as in Study 1, except that the correlation parameter was replaced by $\rho_x = 0.5$. The error $\boldsymbol{\varepsilon}_i$ was generated as in Study 1, with exchangeable correlation $\rho_\varepsilon = 0.2, 0.5$, or 0.8.

For the partial OLS and the partial SIR approaches, we applied the same procedure as in Study 1. For the proposed method, we used a power transformation to recover basis vectors for the central subspace: let $h_j(y_{it}) = y_{it}^j, j = 1, \ldots, s$. Here we set $s = 2$. In an unreported simulation study, we found that increasing s does not make much difference for central subspace estimation. Alternative transformation methods provided in Section 3.1 can also be applied here.

Tables 2, 3, and 4 list the distance under the configurations $(n, p) = (100, 8)$ and $(n, p) = (300, 15)$ for Models I, II, and III. Evidently, the proposed QIF methods are better than the partial OLS and the partial SIR, and the QIF assuming exchangeable correlation is the best. When the correlation of responses

Table 3. Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}'-\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model II from 1,000 simulations.

	Model II	$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
		$n = 100$, $2p = 16$		
partial OLS	Independent	1.3222(0.1456)	1.3406(0.1400)	1.3517(0.1393)
partial SIR	Independent	1.2555(0.2078)	1.2829(0.2045)	1.3060(0.1949)
QIF	Independent	0.8430(0.1255)	0.8975(0.1477)	0.9602(0.1687)
	AR-1	0.8015(0.1548)	0.7789(0.1600)	0.7233(0.1602)
	Exchangeable	0.7917(0.1555)	0.7787(0.1649)	0.7367(0.1665)
		$n = 300$, $2p = 30$		
partial OLS	Independent	1.3657(0.0906)	1.3772(0.0907)	1.3874(0.0926)
partial SIR	Independent	1.1809(0.2060)	1.1967(0.2022)	1.2212(0.1990)
QIF	Independent	0.5789(0.0720)	0.6430(0.0815)	0.7152(0.0919)
	AR-1	0.5757(0.0832)	0.5706(0.0830)	0.5357(0.0778)
	Exchangeable	0.5446(0.0755)	0.5414(0.0750)	0.5096(0.0716)

Table 4. Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}'-\mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model III from 1,000 simulations.

	Model III	$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
		$n = 100$, $2p = 16$		
partial OLS	Independent	1.1462(0.2257)	1.1358(0.2271)	1.1130(0.2359)
partial SIR	Independent	1.2368(0.1883)	1.2293(0.1921)	1.2363(0.1911)
QIF	Independent	0.9571(0.0821)	0.9638(0.0959)	0.9713(0.1089)
	AR-1	0.8141(0.1584)	0.8038(0.1503)	0.7941(0.1385)
	Exchangeable	0.7713(0.1303)	0.7735(0.1247)	0.7755(0.1167)
		$n = 300$, $2p = 30$		
partial OLS	Independent	1.3290(0.0929)	1.3258(0.0947)	1.3205(0.0994)
partial SIR	Independent	1.3363(0.0922)	1.3307(0.0972)	1.3319(0.0978)
QIF	Independent	0.8916(0.0750)	0.9093(0.0893)	0.9272(0.0998)
	AR-1	0.9644(0.1244)	0.9466(0.1222)	0.9269(0.1162)
	Exchangeable	0.6716(0.1004)	0.6592(0.0918)	0.6437(0.0837)

increases, either through the correlations of covariate $\mathbf{x}_i$ in Model I or through the error $\boldsymbol{\varepsilon}_i$ in Models II and III, the proposed method with exchangeable correlation structure is most accurate, while methods assuming independence structure perform poorly. Meanwhile, the QIF assuming AR-1 structure provides very similar estimation as the one assuming exchangeable correlation, because, although we generate both $\mathbf{x}_i$ and $\boldsymbol{\varepsilon}_i$ using the exchangeable correlation structure the combined correlation structure of $\mathbf{y}_i$ is neither exchangeable nor AR-1, due to the nonlinear relationship of the response and covariates.

In general, the proposed QIF dimension reduction method is still applicable if $T < d$, but the partial OLS is not feasible.

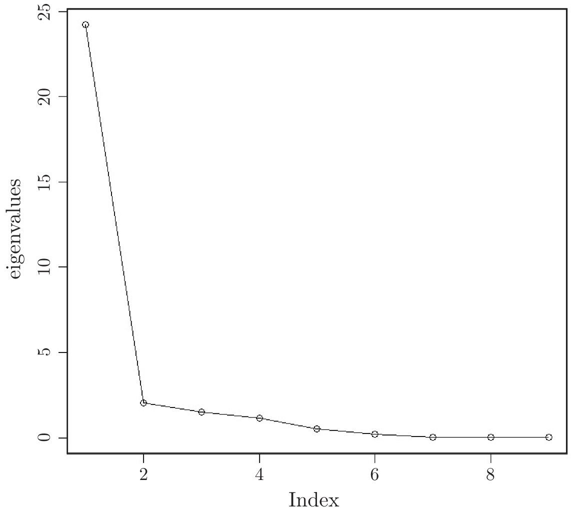

Figure 1. Scree plot of eigenvalues from the partial OLS method for the asthma data by Li and Yin (2009).

6. Asthma Data

We applied the proposed method to an asthma study conducted in Windsor, Ontario, Canada in 1992. This study intends to measure the impact of air pollution on asthmatic patients. The data were originally provided by Professor Paul Corey of the University of Toronto and the Ontario Ministry of Health, and were investigated for model selection in the GEE (Fu (2003)) and partial OLS dimension reduction by Li and Yin (2009). This data set consists of 39 asthmatic patients who were observed on 21 consecutive days. Patients' asthmatic status on difficulty of breathing is recorded as 1 (presence) or 0 (absence) daily, where difficulty of breathing is determined by patients' daily forced expiratory volume. The predictors are daily mean humidity (HUMD), daily mean temperature (TEMP), and seven air pollutants: nitrogen oxide (NO), nitrogen dioxide (NO2), mixture of NO and NO2 (NOX), carbon monoxide (CO), ozone level (OZ), total reduced sulphur (TRS) and coefficient of haze (COH). The data thus contains $n = 39$ patients with cluster size $T = 21$, and dimension of covariates $p = 9$.

We applied the partial OLS by Li and Yin (2009). The scree plot of the eigenvalues of $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)'$ is shown in Figure 1, where each $\boldsymbol{\eta}_t$ is the OLS estimator at each time point $t = 1, \ldots, 21$. To select the structural dimension d, we applied the maximal eigenvalue ratio criterion (Luo, Wang, and Tsai (2009)) discussed in Section 4.1, and $\hat{d} = 1$ was selected. The choice of d can also be observed directly by examing the scree plot in Figure 1, where a sharp drop occurs right after the largest eigenvalue. The corresponding eigenvector associated with the maximum eigenvalue is $\hat{\boldsymbol{\beta}}_{OLS} = (-0.0012, 0.4303, -0.8608, 0.2503, -0.0596,$

Table 5. Logistic regression slope estimates, standard errors and p-values for each $\hat{\boldsymbol{\beta}}$ for the asthma data.

	partial OLS	QIF independent	QIF AR-1	QIF exchangeable
Estimate	-1.6884	0.6879	-0.2307	0.6111
Std. Error	0.9654	0.2002	0.6774	0.1991
p-value	0.0839	0.0006	0.7330	0.0021

$-0.0015, -0.0804,\ 0.0264, -0.0220)'$, and therefore Span$(\hat{\boldsymbol{\beta}}_{OLS})$ is an estimated central subspace. We also observe that the mean sample correlation of the intra-cluster correlation matrix for the responses is 0.6992, and pair correlations among the 21 measurements are quite similar, suggesting a non-negligible exchangeable correlation structure.

For the proposed method, we took $\hat{\boldsymbol{\beta}}_{OLS}$ as an initial value and $\hat{d} = 1$, then calculated the basis for the central subspace using the proposed QIF dimension reduction approach. We employed the QIF assuming the exchangeable, AR-1, and independence correlation structures. The estimated results were:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{Indep} &= (-0.0665, -0.0058, -0.0046, -0.0331, -0.0254, -0.0243, 0.0725,\\
&\quad -0.0071, 0.0513)';\\
\hat{\boldsymbol{\beta}}_{Ar1} &= (0.0247, -0.0020, -0.0478, -0.0160, 0.0187, -0.0149, 0.0552,\\
&\quad -0.0635, 0.0019)';\\
\hat{\boldsymbol{\beta}}_{Exch} &= (-0.0954, -0.0047, 0.0341, -0.0199, -0.0111, 0.0031, 0.0765,\\
&\quad -0.0041, -0.0273)'.
\end{aligned}$$

These differ from the partial OLS estimate. For instance, the angle between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{Exch}$ is 71.82°, indicating a weak correlation between these two estimators.

We conducted logistic regressions of y_{it} given $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ to investigate which method provides the best prediction, where $\hat{\boldsymbol{\beta}}$ is the estimator based on the partial OLS, or the QIF assuming exchangeable, AR-1, or independent correlation, respectively. Table 5 provides the estimators, standard errors, and p-values for the slope of each regression. The QIF dimension reduction with independent and exchangeable correlation structures fits the data better than the other approaches.

The QIF assuming the exchangeable structure is the most accurate, with the smallest MSE compared to other three methods. At each level of the continuous explanatory variable $\mathbf{x}_{it}$, there is only one observation of the response, so the log-odds of receiving $y_{it} = 1$ at each level of $\mathbf{x}_{it}$ is usually infinity or negative infinity. To pool information of adjacent $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$, we divided the range of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ into K intervals of equal length based on the distribution of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$, where $K = 25, 26, 30$, or 25 was applied to each method, respectively. We then calculated

Table 6. MSEs and correlations of four models between the log-odds and the predicted log-odds for the asthma data.

	partial OLS	QIF independent	QIF AR-1	QIF exchangeable
absolute value of correlation	0.4474	0.3051	0.1039	0.5918
MSE	0.2013	0.5603	0.2361	0.2105

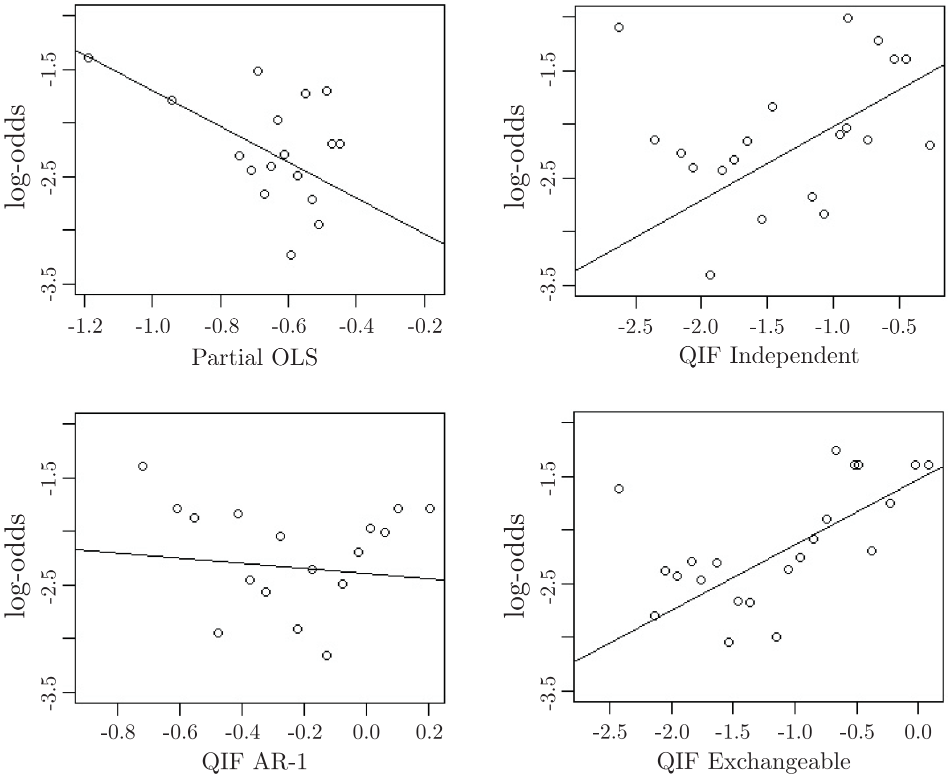

Figure 2. Scatterplots and regression lines after grouping, given by four different methods for the asthma data.

$\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_k$, the average of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ for the k-th interval, and $\mathrm{logit}(\bar{y}_k)$, the log-odds of $\bar{y}_k$, where $\bar{y}_k$ is the average of y_{it} corresponding to $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ in the k-th interval, $k = 1, \ldots, K$. Table 6 lists the correlation between $(\mathrm{logit}(\bar{y}_1), \ldots, \mathrm{logit}(\bar{y}_K))$ and $(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$ and the MSE of $(\alpha_0 + \alpha_1\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \alpha_0 + \alpha_1\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$, where α_0 and α_1 are the logistic regression coefficients. The QIF method assuming exchangeable correlation structure achieves the highest magnitude of regression correlation with a smaller MSE, compared with the QIF assuming independence structure.

Scatterplots of $(\mathrm{logit}(\bar{y}_1), \ldots, \mathrm{logit}(\bar{y}_K))$ against $(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$ for each method are provided in Figure 2. The slope of the partial OLS method (upper-left panel) are very sensitive to the two influential points on the top left, leading to a potentially unstable estimator; while the QIF assuming the exchangeable correlation structure provides a better fitted regression line overall.

7. Discussion

We have addressed the sufficient dimension reduction problem for longitudinal data, with the goal of showing that incorporating intracluster correlation information can achieve more efficiency than assuming independence in both parameter and central subspace estimations. We used the quadratic inference function to incorporate correlation structures and a transformation method to formulate basis vectors for the central subspace. These basis vectors were shown to be consistent and more efficient than estimators assuming independence. The proposed method achieves an overall efficiency for central subspace estimation through combining each efficient estimator of an individual basis vector. Our simulation studies show that the proposed method is quite effective for both binary and continuous data for small and large sample sizes, compared with existing approaches that do not take intracluster correlation into consideration.

Simulation show that even if the correlation structure is misspecified, the efficiency of the proposed estimator is higher than the one assuming independence; our method is quite robust under a small sample size, due to utilizing the entire cluster information for dimension reduction. The proposed method is able to recover the central subspace even when the cluster size is small, it can handle unbalanced data, and is computationally efficient when the cluster size is large.

Further investigation is needed regarding a tuning procedure to select the number of transformations s by minimizing the distance between γ_s and $\text{Span}(\gamma_1, \ldots, \gamma_{s-1})$, along with a penalty function. Another possible research direction is sufficient dimension reduction for binary data (or data with finite support) when the structural dimension is greater than 1. Binary sufficient dimension reduction is quite challenging, since the binary response is invariant for most types of transformation methods. Pooling similar covariate information together so that the log-odds of $Y = 1$ have sufficient variability for estimation is a possible approach.

Supplementary Materials

The proofs of the theorems and the lemma in this paper are given in the online supplemental material available at `http://www3.stat.sinica.edu.tw/statistica/`.

Acknowledgement

The authors are very grateful to the Co-Editor, two referees and an associate editor for their insightful comments and suggestions that have significantly improved the manuscript. The authors are also grateful to Professor Lexin Li for stimulating discussion on this topic. The research was supported by a National Science Foundation Grant (DMS-0906660 and DMS-1308227).

References

Beckenback, E. F. and Bellman, R. (1965). *Inequalities.* Springer-Verlag, Berlin.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91**, 983-992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100**, 410-428.

Cook, R. D. and Ni, L. (2006). Using intra-slice covariances for improved estimation of the central subspace in regression. *Biometrika* **93**, 65-74.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction," by K.C. Li. *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. and Weisberg, S. (1994) Transforming a response variable for linearity. *Biometrika* **81**, 731-738.

Cook, R. D. and Yin, X. (2001). Dimension-reduction and visualization in discriminant analysis. *Austral. N. Z. J. Statist.* **43**, 147-200.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika* **97**, 279-294.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics* **59**, 126-132.

Fung, W. K., He, X., Liu, L. and Shi, P. D. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12**, 1093-1114.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Li, B., Cook, R. D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Ann. Statist.* **31**, 1636-1668.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37**, 1272-1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025-1039.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.

Li, L. and Yin, X. (2009). Longitudinal data analysis using sufficient dimension reduction method. *Comput. Statist. Data Anal.* **53**, 4106-4115.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Luo, R., Wang, H. and Tsai, C. L. (2009). Contour projected dimension reduction. *Ann. Statist.* **37**, 3743-3778.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107**, 168-179.

Ma, Y. and Zhu, L. (2013a). A review on dimension reduction. *Internat. Statist. Rev.* **81**, 134-150.

Ma, Y. and Zhu, L. (2013b). Efficiency loss and the linearity condition in dimension reduction. *Biometrika* **100**, 371-383.

Ma, Y. and Zhu, L. (2013c).. Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**, 250-268.

Ma, Y. and Zhu, L. (2013d).. On estimation efficiency of the central mean subspace. *J. Roy. Statist. Soc. Ser. B* (in press).

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59-67.

Pardoe, I., Yin, X. and Cook, R. D. (2007). Graphical tools for quadratic discriminant analysis. *Technometrics* **49**, 172-183.

Pfeiffer, R. M., Forzani, L. and Bura, E. (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statist. Medicine* **31**, 2414-2427.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika* **87**, 823-836.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811-821.

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

Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98**, 968-979.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. *J. Roy. Statist. Soc. Ser. B* **64**, 159-176.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39**, 3392-3416.

Zhou, J. (2009). Robust dimension reduction based on canonical correlation. *J. Multivariate Anal.* **100**, 195-209.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649-1668.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *J. Amer. Statist. Assoc.* **107**, 701-710.

Zhu, L. P. and Zhu, L. X. (2009). Dimension reduction for conditional variance in regressions. *Statist. Sinica* **19**, 869-883.

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: xuanbi2@illinois.edu

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: anniequ@illinois.edu

(Received June 2013; accepted May 2014)