| Statistica Sinica Preprint No: SS-2017-0347 | |
|---|---|
| **Title** | Understanding and Utilizing the Linearity Condition in Dimension Reduction |
| **Manuscript ID** | SS-2017-0347 |
| **URL** | http://www.stat.sinica.edu.tw/statistica/ |
| **DOI** | 10.5705/ss.202017.0347 |
| **Complete List of Authors** | Yanyuan Ma<br>Fei Jiang and<br>Masayuki Henmi |
| **Corresponding Author** | Yanyuan Ma |
| **E-mail** | yanyuanma@gmail.com |

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Statistica Sinica

# Understanding and Utilizing the Linearity Condition in Dimension Reduction

Yanyuan Ma, Fei Jiang, Masayuki Henmi

*Penn State University, University of Hong Kong, Institute of Statistical Mathematics*

*Abstract:* When using inverse regression methods in dimension reduction models, the popular linearity condition has a paradoxical effect: ignoring the linearity condition yields a more efficient estimator than making use of the linearity condition. By considering classes of parametric models, which include the linearity condition as a special case, we examine this phenomenon using a geometry approach, and provide an intuitive and extended explanation. Our findings explain what the real cause of the paradox is, indicate how to properly handle the linearity condition and reveal the true role of the linearity condition. Our analysis directly leads to new estimators that further improve the existing efficient estimator that did not specifically account for the linearity condition and the possible constant variance

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

condition.

# 1. Introduction

The linearity condition, often jointly with the constant variance condition, is a popular assumption when data require dimension reduction (Li and Duan, 1989; Li, 1991; Cook and Weinberg, 1991; Li, 1992; Cook, 1998; Cook and Li, 2002; Li and Wang, 2007). These conditions are routinely assumed in the dimension reduction literature, where the central model assumption is that the response $Y$ is linked to the covariates $\mathbf{X}$ via some linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, where $\boldsymbol{\beta}$ is a matrix; that is, $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The linearity condition states that the covariate vector $\mathbf{X}$ satisfies that $E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The constant variance condition states that $\mathrm{cov}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is a constant matrix. However, a puzzling phenomenon is described in the numerical experiment of Ma and Zhu (2012), where it is discovered that ignoring these conditions, even if they indeed hold, yields better results than making use of these conditions. It was later discovered that the gain is in terms of estimation efficiency. More precisely, assuming the linearity condition and, if needed, the constant variance condition to hold, Ma and Zhu (2013a) showed that if we ignore these conditions and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

nonparametrically estimate the relevant quantities, which would be known under these conditions, then the resulting estimation variance of the inverse regression method decreases. For brevity, in the following, we refer to this phenomenon as the linearity puzzle.

Although a mathematical proof is provided, Ma and Zhu (2013a) do not provide an intuitive explanation behind the astonishing phenomenon. This is somewhat a pity because a mathematical proof does not necessarily or at least easily lead to clear understanding of the phenomenon. Thus, the intuition on why linearity puzzle occurs is still missing. On the other hand, a similar paradoxical phenomenon related to missing values is familiar to researchers in statistics. There, it was known that in implementing the inverse probability weighting method, using an estimated weight, even when the true weight is known, can reduce the estimation variability (Hirano et al., 2003). This counterintuitive phenomenon was later beautifully explained using information geometry by Henmi and Eguchi (2004), thus revealing the underlying structure of the inverse probability weighting estimator and the inherent reason for the paradox.

This motivates us to use the tool of information geometry to inspect the linearity puzzle and to provide an intuitive explanation in a style similar to

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

that of Henmi and Eguchi (2004). This plan turns out to be partially feasible. Although we managed to understand and explain the intuition behind the linearity puzzle, our approach is a mixture of information geometry and algebra; thus it is quite different to that of Henmi and Eguchi (2004). Information geometry (Amari and Kawanabe, 1997; Amari and Nagaoka, 2007) usually employs differential geometry tools; here, we adopt it in a wider sense. In the following, we first describe the linearity puzzle in its general form. Then, we inspect a series of models that bridge the completely known function specified in the linearity condition and a completely unknown function when this condition is given up in combination with various estimation procedures. By investigating this more general model setting, we are able to embed the linearity puzzle within a bigger picture, and gain an intuitive understanding of why and how it occurs.

We expand our understanding of these phenomena by investigating whether the linearity and constant variance conditions can be used smartly and "properly" so that they contribute to the estimation in a positive way. We answer this question by deriving the semiparametric efficiency bound under either and both conditions. Our results indicate that if these conditions are accounted for appropriately, they can indeed contribute to bene-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

fitting the estimation in terms of reducing the estimation variability. Encouraged by this discovery, we further devise two new estimators. The first is optimal under the linearity condition, and the second is optimal under both linearity and constant variance conditions. We illustrate the advantages of the new estimators over the efficient estimator proposed in Ma and Zhu (2013b), both theoretically and numerically, by means of a series of simulation studies.

## 2. Linearity puzzle

Let $\mathbf{X}$ be a $p \times 1$ random vector, and let $Y$ be a univariate random response. In the dimension reduction literature, the goal of estimating the column space of $\boldsymbol{\beta}$ can be expressed equivalently as estimating the lower $(p-d) \times d$ block of the $p \times d$ matrix $\boldsymbol{\beta}$, while fixing the upper $d \times d$ matrix as the identity. A very general class of estimators can be constructed from the estimating function $\mathbf{g}(Y)[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}}$, where $\mathbf{g}$ is a prespecified length $p_g$ vector function, $\mathbf{a}$ is a prespecified length $p_a$ vector function, and $p_g p_a = (p-d)d$. Ma and Zhu (2013a) showed that the estimator of $\boldsymbol{\beta}$ obtained

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

from solving

$$\sum_{i=1}^{n} \mathbf{g}(Y_i)\{\mathbf{a}(\mathbf{X}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^{\mathrm{T}} = \mathbf{0} \tag{21}$$

is less efficient than the estimator of $\boldsymbol{\beta}$ obtained from solving

$$\sum_{i=1}^{n} \mathbf{g}(Y_i)\{\mathbf{a}(\mathbf{X}_i) - \widehat{\mathbf{m}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^{\mathrm{T}} = \mathbf{0}. \tag{22}$$

Here, $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \equiv E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$ is the true expectation of $\mathbf{a}(\mathbf{X})$ conditional on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, and $\widehat{\mathbf{m}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \equiv \widehat{E}\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$ is a kernel-based nonparametric estimator of $E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$. Here and throughout the text, when we say one estimator (say estimator $\widetilde{\boldsymbol{\beta}}$) is less efficient than the other (say estimator $\widehat{\boldsymbol{\beta}}$), we mean that the difference of the asymptotic variances of the two estimators ($\mathrm{cov}(\widetilde{\boldsymbol{\beta}}) - \mathrm{cov}(\widehat{\boldsymbol{\beta}})$) is positive-definite. This result is quite counterintuitive because usually one would expect the additional estimation to inflate the overall estimation variability.

# 3. General parametric model case

To gain a comprehensive view of the linearity puzzle described in Section 2, and to generalize it to all parametric models, we consider estimating equations of the form

$$\sum_{i=1}^{n} \mathbf{g}(Y_i)[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta})\}]^{\mathrm{T}} = \mathbf{0}. \tag{33}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Here, $\boldsymbol{\alpha} \in \mathbb{R}^{p_\alpha}$ is a parameter, $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha})$ is a true parametric model of $E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$, i.e. there exists $\boldsymbol{\alpha}_0(\boldsymbol{\beta})$ such that $\mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\} = E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$ for any $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}$. We can view Li and Dong (2009) as a special case of this consideration. Because we substitute in the conditional mean of the covariate function $\mathbf{a}(\mathbf{X})$, we call the estimators derived from (33) the family of plug-in estimators. We can view (33) as a transition from the completely known expectation in (21) to the completely unknown expectation in (22). In (33), when $p_\alpha = 0$, we obtain (21), and when $p_\alpha = \infty$, we obtain (22). Various methods of estimating $\boldsymbol{\alpha}$ exist, based on the regression model

$$\mathbf{a}(\mathbf{X}) = \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}) + \boldsymbol{\epsilon},$$

where $E(\boldsymbol{\epsilon} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{0}$. In general, an estimator $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ can be written as the root of

$$\sum_{i=1}^{n} \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\{\mathbf{a}(\mathbf{X}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha})\} = \mathbf{0}, \tag{34}$$

where $\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ is a $p_\alpha \times p_a$ matrix. Obviously, different choices of $\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ lead to different weighted least squares (WLS) estimators.

To compare the performance of the estimators obtained from solving various estimating equations, we use $\widetilde{\boldsymbol{\beta}}$ to denote the estimator from (21),

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

$\widehat{\boldsymbol{\beta}}$ to denote that from (22), and $\breve{\boldsymbol{\beta}}$ to denote that from (33). The relative performance of the three estimators is summarized in Theorem 1. Here and throughout the text, we use vec($\mathbf{A}$) to denote the vector formed by the columns of the matrix $\mathbf{A}$, and we use vecl($\mathbf{A}$) to denote vec($\mathbf{A}_L$), where $\mathbf{A}_L$ is the lower submatrix of $\mathbf{A}$, excluding the upper square submatrix. Further, define

$$\begin{aligned}
\boldsymbol{\Sigma}_A &= E\left([\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}] \otimes \frac{\partial E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}}{\partial \mathrm{vecl}(\boldsymbol{\beta})^{\mathrm{T}}}\right), \\
\mathbf{B}_1 &= E\left[\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{m}_{\boldsymbol{\alpha}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right], \\
\mathbf{B}_2 &= E\left[\mathbf{m}_{\boldsymbol{\alpha}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\} \otimes \mathbf{g}(Y)\right],
\end{aligned}$$

where $\mathbf{m}_{\boldsymbol{\alpha}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\} = \partial \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha})/\partial \boldsymbol{\alpha}^{\mathrm{T}}|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0(\boldsymbol{\beta})}$, and $\otimes$ denotes the Kronecker product.

**Theorem 1.** *Under the conditions given in the Supplementary Material, the estimators $\widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ satisfy*

$$\begin{aligned}
&\sqrt{n}\boldsymbol{\Sigma}_A vecl(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] + o_p(1), \\
&\sqrt{n}\boldsymbol{\Sigma}_A vecl(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \;& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

$$
\begin{aligned}
& +o_p(1), \\
& \sqrt{n}\boldsymbol{\Sigma}_A vecl(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right] \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\} \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] + o_p(1).
\end{aligned}
$$

The results for $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ are obtained in Ma and Zhu (2013a); we provide the derivation of the result for $\breve{\boldsymbol{\beta}}$ in the Supplement Material **??**. We show the relative performance of the estimators $\widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ through their influence functions in the left panel of Figure 1,

Clearly, $\widehat{\boldsymbol{\beta}}$ has the smallest variance because $\mathrm{vec}([\mathbf{g}(Y) - E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}][\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ is orthogonal with both $\mathrm{vec}(E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ and $\mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$. Thus, we mainly illustrate the gains of $\widetilde{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ over $\widehat{\boldsymbol{\beta}}$. In Figure 1, we set the origin at $\mathrm{vec}([\mathbf{g}(Y) - E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}][\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$, which can be understood as a vector orthogonal to the plane plotted in Figure 1. We set the center of the circle at $\mathrm{vec}(E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$. Thus, the circle contains all functions for which the difference between the function and $\mathrm{vec}(E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ has

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

the same length as $\text{vec}(E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ itself. We highlight three different scenarios for $\mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$, indicating three possible outcomes of the relative performance of $\breve{\boldsymbol{\beta}}$, namely falling inside, outside, and on the circle. As we can see, depending on the choice of $\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ and the effect of the parametric model $\mathbf{m}_{\boldsymbol{\alpha}}$ through $\mathbf{B}_1$ and $\mathbf{B}_2$, $\breve{\boldsymbol{\beta}}$ could have larger variance than $\widetilde{\boldsymbol{\beta}}$, smaller variance than $\widetilde{\boldsymbol{\beta}}$ or the same variance as $\widetilde{\boldsymbol{\beta}}$, illustrated as the vector a, b and c to the center of the circle respectively in the left panel of Figure 1.

In addition to all these different relative performances between $\breve{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, the relative performances between different parametric models are also very complex. Generally speaking, there exists no monotonicity result among different parametric models, even among nested parametric models. This is reflected in the left panel of Figure 1, in that the point corresponding to $\mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$ can be anywhere in the figure, and its distance to the center of the circle can be as small as zero or as large as one wishes.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

# 4. General parametric model with constant variance and OWLS

We now direct our attention to a smaller class of estimators than those described in (34). The constant variance condition requires $\text{var}(\boldsymbol{\epsilon} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ to be a constant matrix, denoted as $\mathbf{Q}$. However, in general, the error variance can depend on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ as well; therefore we write the variance-covariance matrix more explicitly as $\mathbf{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. It is well known that among all WLS estimators for regression models, the optimal weighted least squares (OWLS) estimator is optimal, in that it has the smallest estimation variance of the various WLS estimators. The optimal weight matrix is $\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. Thus, we now consider exclusively the OWLS estimator $\widehat{\boldsymbol{\alpha}}$, which solves (34), with $\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ replaced by $\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha})\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. The estimating equation corresponding to (34) is then

$$\sum_{i=1}^{n} \mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha})\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\{\mathbf{a}(\mathbf{X}_i) - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha})\} = \mathbf{0}. \tag{45}$$

Define

$$\mathbf{B}_3 = E\left[\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{m}_{\boldsymbol{\alpha}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right],$$

which a special case of $\mathbf{B}_1$. The general result concerning $\breve{\boldsymbol{\beta}}$ in Theorem 1 subsequently reduces to a special structure that has a unique orthogonality

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

property. We explicitly describe the properties of the OWLS estimator in Theorem 2. The OWLS estimator is slightly different from the estimators that solve (34) because $\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ in (34) does not contain $\boldsymbol{\alpha}$. We give the proof of Theorem 2 in the Supplementary Material **??**.

**Theorem 2.** *When the OWLS estimator $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ is obtained by solving (45), the resulting estimator $\breve{\boldsymbol{\beta}}$ satisfies*

$$
\begin{aligned}
& \sqrt{n}\boldsymbol{\Sigma}_A vecl(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
= \; & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right] \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\} \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right] \\
& + o_p(1).
\end{aligned}
$$

*In addition,*

$$
\mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \left[\mathbf{a}(\mathbf{X}) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]
$$

*is orthogonal to*

$$
\begin{aligned}
& \left[\mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} - \mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right] \\
& \times \left[\mathbf{a}(\mathbf{X}) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right].
\end{aligned}
$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Similarly to Section 3, we inspect the results from Theorem 2 from an information geometry point of view in the right panel of Figure 1 to gain an intuitive understanding of the results. Now, because of the orthogonality property established in Theorem 2, the length of

$$\begin{aligned}&\left[\mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} - \mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right] \\ &\times \left[\mathbf{a}(\mathbf{X}) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]\end{aligned}$$

is the distance from the center of the circle to an arbitrary line that goes through the origin; hence, it is always shorter than the radius itself, which is the length of

$$\mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}\left[\mathbf{a}(\mathbf{X}) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right].$$

It is now immediately clear that the variance of the estimator $\check{\boldsymbol{\beta}}$ based on any parametric model $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha})$ with OWLS estimation is always smaller than or equal to the variance of the estimator $\widetilde{\boldsymbol{\beta}}$ based on the known $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. It is also clear that the best parametric model must satisfy $\mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$, for example, by any parametric model that satisfies $\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\} = \mathbf{I}_{p_a} \otimes E\{\mathbf{g}(Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}\mathbf{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$, which would reduce the estimation variance to the minimum, i.e. as small as the variance of $\widehat{\boldsymbol{\beta}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

We now take a closer look at the parametric models to see how different sizes of various parametric models affect the variances of the resulting $\breve{\boldsymbol{\beta}}$'s. In fact, Figure 1 is simplified. For example, the vector from the origin to a, b or c is $\mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$ when it is a two-dimensional vector. When the number of parameters $p_\alpha$ is larger, we should use a plane, a hyperplane, etc. that goes through the origin to represent it. When the parametric models are nested, we subsequently will obtain a plane that contains the line, a hyperplane that contains the plane, and so on. Correspondingly, the circle becomes a sphere, a hypersphere, and so on. As a result, the distances between the circle center to the line, plane, hyperplane will decrease accordingly. This implies that the variances of $\breve{\boldsymbol{\beta}}$ will decrease as the complexity of the parametric models $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha})$ increase. This general observation, in that in terms of the estimation variability of $\breve{\boldsymbol{\beta}}$, "bigger parametric model is always better than a smaller nested model" naturally suggests that when the model is "maximized", we should obtain the optimal estimation variance of $\breve{\boldsymbol{\beta}}$. Intuitively, we would consider a nonparametric model as the "maximum" parametric model, hence this would explain why the nonparametric-based estimator $\widehat{\boldsymbol{\beta}}$ has the smallest variance.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

We now explain how to rationalize the intuition that the kernel based nonparametric estimator can be viewed as an extreme case of the OWLS estimator when the number of parameters is sufficiently large to yield a "nonparametric" model. Consider the parametric model where we put a different mean function value at each different $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ value, i.e. $\mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{c}_0$ when $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0$. In this case, to estimate $\mathbf{c}_0$, the corresponding OWLS estimator becomes

$$\sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\{\mathbf{a}(\mathbf{X}_i) - \mathbf{c}_0\} = \mathbf{0},$$

which yields

$$\mathbf{c}_0 = \frac{\sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\mathbf{a}(\mathbf{X}_i)}{\sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)} = \frac{\sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\mathbf{a}(\mathbf{X}_i)}{\sum_{i=1}^{n} I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)}.$$

Here $\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ appears only when $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0$ hence dropped out from both the numerator and the denominator. Of course we would like to further assume smoothness of the mean function, hence we modify the indicator function $I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)$ to $K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)$. This modifies the above display to

$$\mathbf{c}_0 = \frac{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)\mathbf{a}(\mathbf{X}_i)}{\sum_{i=1}^{n} K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_0)},$$

which is exactly the expression of the Nadaraya-Watson nonparametric regression kernel estimator.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

# 5. Role and correct use of the linearity/constant variance conditions

Having obtained the results in Section 4 and understood why the nonparametric estimator is the most efficient choice among all possible models of $E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}$ both mathematically and intuitively, we are now in the position to take a closer look at the popular linearity condition and the constant variance condition to see the benefits and cost these conditions bring. This turns out to be not a trivial task at all. In order to proceed properly, we take the approach of Bickel et al. (1993); Tsiatis (2006) and derive the nuisance tangent space, its orthogonal complement, and the efficient score, for the case with the linearity condition and the case with both linearity and constant variance conditions.

We first investigate the case when only the linearity condition is assumed to hold. We can express the likelihood of one typical observation $(\mathbf{X}, Y)$ as

$$f(\mathbf{X}, Y) = f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) f_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2) f_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y),$$

where $\boldsymbol{\epsilon}_2 = \mathbf{X}_2 - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) = \mathbf{X}_2 - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\beta} = (\mathbf{I}_d, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. We use $f_1, f_2$, and $f_3$ to denote the probability density

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

function (pdf) of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, the pdf of $\boldsymbol{\epsilon}_2$ conditional on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and the pdf of $Y$ conditional on $\mathbf{X}$, which by the model assumption, is a function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and $Y$ only. We write $\mathbf{Q}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = E(\boldsymbol{\epsilon}_2^{\otimes 2} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$.

**Theorem 3.** *Assume that the linearity condition holds. Then, when estimating $\boldsymbol{\beta}_2$, the nuisance tangent space is $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$, where*

$$\begin{aligned}
\Lambda_1 &= \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} < \infty, \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d}\right], \\
\Lambda_2 &= \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\boldsymbol{\epsilon}_2\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0},\right. \\
&\qquad \left. E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2)\} < \infty, \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2) \in \mathcal{R}^{(p-d)d}\right], \\
\Lambda_3 &= \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\} < \infty,\right. \\
&\qquad \left. \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d}\right].
\end{aligned}$$

*The nuisance tangent space orthogonal complement is*

$$\begin{aligned}
\Lambda^{\perp} &= \left[\mathbf{g}(\mathbf{X}, Y) : E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y\} = \mathbf{0}, E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\epsilon}_2\} = \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\boldsymbol{\epsilon}_2,\right. \\
&\qquad \left. E\{\mathbf{g}^{\mathrm{T}}(\mathbf{X}, Y)\mathbf{g}(\mathbf{X}, Y)\} < \infty, \mathbf{g}(\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d}, \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d\times(p-d)}\right].
\end{aligned}$$

*The efficient score function is*

$$\begin{aligned}
&\mathbf{S}_{\text{eff}}(\mathbf{X}, Y, \boldsymbol{\beta}) \\
&= \operatorname{vec}\left(\boldsymbol{\epsilon}_2\frac{\partial \log f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} + \frac{\partial \mathbf{Q}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left[\mathbf{I}_d \otimes \left\{\mathbf{Q}_2^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\boldsymbol{\epsilon}_2\right\}\right] + \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\boldsymbol{\epsilon}_2^{\mathrm{T}}\right. \\
&\quad \left.\times\mathbf{Q}_2^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\frac{\partial \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} + \boldsymbol{\epsilon}_2\frac{\partial \log f_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right) + \frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)}{\partial \operatorname{vec}(\boldsymbol{\beta}_2)}\mathbf{Q}_2^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\boldsymbol{\epsilon}_2.
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Despite of the complexity of the analytic form of the efficient score described in Theorem 3, it immediately reveals that the linearity condition indeed has contribution towards estimation efficiency. This is because the efficient score under this condition is very different from that without this condition, as given in Ma and Zhu (2013b). Thus, to take full advantage of this condition and to achieve optimal efficiency require estimating $\mathbf{Q}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$, $f_1, f_3$ and their derivatives with respect to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The proof of Theorem 3 is provided in the Supplementary Material **??**.

Under the additional constant variance condition, the variance of $\boldsymbol{\epsilon}_2$ satisfies $E(\boldsymbol{\epsilon}_2^{\otimes 2} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{Q}_2 = \mathbf{I} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}$, which does not vary with $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The analysis of the efficient estimation under both the linearity and constant variance conditions follows the same spirit, but is even more tedious and technical. In this case, we define $\mathbf{Q}_2(\boldsymbol{\beta}_2) = E[\{\mathbf{X}_2 - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\}^{\otimes 2} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}]$, which does not depend on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ according to the assumption, and we assume $\mathbf{Q}_2(\boldsymbol{\beta}_2) = \mathbf{D}^{\otimes 2}(\boldsymbol{\beta}_2)$. We further define $\widetilde{\boldsymbol{\epsilon}}_2 = \mathbf{D}^{-1}(\boldsymbol{\beta}_2)\{\mathbf{X}_2 - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\}$, and write the pdf of $(\mathbf{X}, Y)$ as

$$f(\mathbf{X}, Y) = f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \det\{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\} f_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) f_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y),$$

where $f_1$ and $f_3$ are as before, while $f_2$ is now subject to the mean zero, variance identity condition. We present the results in Theorem 4 and provide

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

a sketch of the proof in the Supplementary Material **??**.

**Theorem 4.** *Assume the linearity and constant variance conditions hold. Then, when estimating $\boldsymbol{\beta}_2$, the nuisance tangent space is $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$, where*

$$\begin{aligned}
\Lambda_1 &= \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} < \infty, \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d}\right] \\
\Lambda_2 &= \Big[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\widetilde{\boldsymbol{\epsilon}}_2\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, \\
&\quad E\{\mathrm{vec}(\widetilde{\boldsymbol{\epsilon}}_2\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}})\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)\} < \infty, \\
&\quad \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \in \mathcal{R}^{(p-d)d}\Big] \\
\Lambda_3 &= \Big[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\} < \infty, \\
&\quad \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d}\Big].
\end{aligned}$$

*The nuisance tangent space orthogonal complement is*

$$\begin{aligned}
\Lambda^{\perp} = \Big[&\mathbf{g}(\mathbf{X}, Y) : E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2\} = \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\widetilde{\boldsymbol{\epsilon}}_2 + \mathbf{B}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathrm{vec}(\widetilde{\boldsymbol{\epsilon}}_2\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} - \mathbf{I}_{p-d}), \\
&E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y\} = \mathbf{0}, E\{\mathbf{g}^{\mathrm{T}}(\mathbf{X}, Y)\mathbf{g}(\mathbf{X}, Y)\} < \infty, \mathbf{g}(\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d}\Big].
\end{aligned}$$

*The efficient score function is*

$$\begin{aligned}
\mathbf{S}_{\mathrm{eff}}(\mathbf{X}, Y, \boldsymbol{\beta}) = &\ \mathrm{vec}\left(\mathbf{D}(\boldsymbol{\beta}_2)\widetilde{\boldsymbol{\epsilon}}_2\frac{\partial \log f_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}} + \mathbf{D}(\boldsymbol{\beta}_2)\widetilde{\boldsymbol{\epsilon}}_2\frac{\partial \log f_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)}{\partial \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\right) \\
&-\mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\widetilde{\boldsymbol{\epsilon}}_2 + \mathbf{K}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\mathbf{v} - \mathbf{K}_4(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)\mathbf{v}.
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

*Here,* $\mathbf{v} = \mathbf{v}_1 - E(\mathbf{v}_1\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\widetilde{\boldsymbol{\epsilon}}_2$, $\mathbf{v}_1 = \mathrm{vec}(\widetilde{\boldsymbol{\epsilon}}_2\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} - \mathbf{I}_{p-d})$,

$$
\begin{aligned}
\mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) &= -\left\{\frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}} \otimes \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) + \frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)}\right\}\{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}} \\
\mathbf{K}_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) &= \mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)E(\mathbf{v}_1\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\left\{E(\mathbf{v}\mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\right\}^{-1} \\
\mathbf{C}_1(\boldsymbol{\beta}_2) &= \left[\mathrm{vec}\left\{\mathbf{D}^{\mathrm{T}}(\boldsymbol{\beta}_2)\frac{\partial\{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}}}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)_1}\right\}, \ldots, \mathrm{vec}\left\{\mathbf{D}^{\mathrm{T}}(\boldsymbol{\beta}_2)\frac{\partial\{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}}}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)_{(p-d)d}}\right\}\right]^{\mathrm{T}} \\
\mathbf{K}_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) &= \mathbf{C}_1(\boldsymbol{\beta}_2) - \left[\frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}}\left\{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\right\}^{\mathrm{T}}\right] \otimes \mathbf{D}(\boldsymbol{\beta}_2) \\
\mathbf{K}_4(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) &= -\mathbf{K}_3(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2)E\left[\mathrm{vec}\left\{\widetilde{\boldsymbol{\epsilon}}_2\frac{\partial \log f_2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)}{\partial \widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}}}\right\}\mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\right]\{E(\mathbf{v}\mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\}^{-1}.
\end{aligned}
$$

Similar observations can be made regarding the efficient estimator under both the linearity condition and the constant variance condition. Theoretically, the additional constant variance condition indeed further improves the optimal efficiency bound over that under linearity condition only. This conclusion is drawn directly based on the fact that the two efficient score functions given in Theorems 3 and 4 are different (Bickel et al., 1993; Tsiatis, 2006).

# 6. Numerical evaluation

We perform a series of numerical experiments to evaluate the efficient estimators. We simulate $p - d$ independent variables $u_j$ $(j = 1, \ldots, p - d)$ from a Weibull distribution with shape parameter one and scale parameter

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

two. Let $\boldsymbol{\epsilon}_2 = \mathbf{v}^{-1/2}(\mathbf{u} - \mathbf{m})$, where $\mathbf{u} = (u_1, \ldots, u_{p-d})^{\mathrm{T}}$, and $\mathbf{m}$ and $\mathbf{v}$ are, respectively, the mean and variance of $\mathbf{u}$. Thus, $\boldsymbol{\epsilon}_2$ is a random vector with mean $\mathbf{0}$, and variance-covariance matrix $\mathbf{I}_{p-d}$. We consider $Z$ from two distributions, a logistic distribution and a centered gamma distribution. Both distributions have mean zero and variance $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$. To generate the covariates, we consider three models:

(I) $\mathbf{X}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}Z + (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1/2}|Z|\{\mathbf{I}_{p-d} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$;

(II) $\mathbf{X}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}Z + \{\mathbf{I}_{p-d} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$,

(III) $\widetilde{\mathbf{X}}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}(Z^2 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}) + \{\mathbf{I}_{p-1} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$.

We then let $X_1 = Z - \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}_2$, $\mathbf{X} = (X_1, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$ in models (I) and (II). For model (III), we let $\widetilde{X}_1 = Z - \boldsymbol{\beta}_2^{\mathrm{T}}\widetilde{\mathbf{X}}_2$, and form $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \widetilde{\mathbf{X}}_2^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{X} = \{\mathrm{var}(\widetilde{\mathbf{X}})\}^{-1/2}\widetilde{\mathbf{X}}$, where $\mathrm{var}(\widetilde{\mathbf{X}})$ is the variance-covariance matrix of $\widetilde{\mathbf{X}}$. Our construction is designed under $d = 1$, and ensures that the resulting covariate vector $\mathbf{X}$ satisfies $E(\mathbf{X}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{X}) = \mathbf{I}_p$. In addition, model I satisfies the linearity condition, but not the constant variance condition, whereas model (II) satisfies both the linearity and the constant variance conditions. Specifically, both models satisfy $E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. On the other hand, for model (I), $\mathrm{cov}(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2\{\mathbf{I}_p -$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

$\boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\}$, which is not a constant matrix, whereas for model (II), $\mathrm{cov}(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{I}_p - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$. In contrast to models (I) and (II), model (III) does not satisfy the linearity condition. We subsequently simulate $Y$ from the normal distribution with mean $\sin(2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ and variance $\log\{2 + (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2\}$.

We now implement the three optimal estimators: (a) the efficient estimator given in Ma and Zhu (2013a), which does not make use of linearity or constant variance condition, and is optimal without assuming these relations; (b) the efficient estimator proposed in Theorem 3, which assumes only the linearity condition, and is optimal if this assumption is indeed satisfied; and (c) the efficient estimator proposed in Theorem 4, which assumes both the linearity and the constant variance conditions, and is optimal when both conditions hold. We set $\boldsymbol{\beta} = (1, 0.2, 0.3, 0.4)^{\mathrm{T}}$ and compare the three estimators under the various model assumptions with a sample size of $n = 50$.

We illustrate the performance of the three estimators across 1000 simulations in Tables 1 and 2, which present the estimation absolute biases and standard errors. In addition, we create the boxplots for the squared distances of the estimators to the true parameter values, defined as $\{(p -$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

$d)d\}^{-1}\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_2^2$ in Figure 2 and 3. From Table 1 and2, and Figure 2, we can see that under model (I), where only the linearity condition is satisfied, estimator (b) has smaller estimation error compared with estimator (a). This is because estimator (b) utilizes the linearity condition properly to further improve the estimation efficiency over that of estimator (a). However, estimator (c) goes too far in further assuming the constant variance condition and hence is biased. In contrast, under model (II), where both the linearity and constant variance conditions are satisfied, all estimators are consistent while estimator (c) yields the smallest estimation error. This is because estimator (c) increases the estimation efficiency by further taking into account the additional constant variance condition in an optimal way. Finally, under model (III), where the linearity condition does not hold, estimator (a) performs best, because the other two estimators are both obtained under wrong models and hence are both inconsistent.

We also considered $d = 2$. In this case, we first generate $\mathbf{u}$ and $\boldsymbol{\epsilon}_2$ in the same way as before. To construct $\mathbf{Z}$, we write the $(i, j)$ element of $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ as $\sigma_{ij}$, and first generate $Z_1$ from a centered gamma distribution with mean zero and variance $\sigma_{11}$. Then, we generate $Z_2$ from a logistic distribution with mean $\sigma_{12}\sigma_{11}^{-1}Z_1$ and variance $\sigma_{22} - \sigma_{12}\sigma_{11}^{-1}\sigma_{21}$. This ensures that the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

vector $\mathbf{Z} = (Z_1, Z_2)^{\mathrm{T}}$ has mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$. To generate the covariates, we consider three models, similarly to the $d = 1$ case:

(I) $\mathbf{X}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\mathbf{Z} + \sigma_{11}^{-1/2}|Z_1|\{\mathbf{I}_{p-1} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$;

(II) $\mathbf{X}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\mathbf{Z} + \{\mathbf{I}_{p-1} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$;

(III) $\widetilde{\mathbf{X}}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}) + \{\mathbf{I}_{p-1} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$.

We then let $\mathbf{X}_1 = \mathbf{Z} - \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}_2$, and $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$ in models (I) and (II). For model (III), we let $\widetilde{\mathbf{X}}_1 = \mathbf{Z} - \boldsymbol{\beta}_2^{\mathrm{T}}\widetilde{\mathbf{X}}_2$, $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1^{\mathrm{T}}, \widetilde{\mathbf{X}}_2^{\mathrm{T}})^{\mathrm{T}}$ and form $\mathbf{X} = \{\mathrm{var}(\widetilde{\mathbf{X}})\}^{-1/2}\widetilde{\mathbf{X}}$. Careful algebraic calculations can then be employed to verify that the construction in the three models satisfies $E(\mathbf{X}) = \mathbf{0}$ and $\mathrm{var}(\mathbf{X}) = \mathbf{I}_p$. In addition, $\mathbf{X}$ satisfies only the linearity condition in model (I), but also satisfies the constant variance condition in model (II). It does not satisfy the linearity condition in model (III). We then generate $Y$ from a normal distribution with mean $\sin(2\boldsymbol{\beta}_{\cdot 1}^{\mathrm{T}}\mathbf{X})$ and variance $\log\{2 + (\boldsymbol{\beta}_{\cdot 2}^{\mathrm{T}}\mathbf{X})^2\}$, where $\boldsymbol{\beta}_{\cdot 1}$ and $\boldsymbol{\beta}_{\cdot 2}$ denote the first and second columns of $\boldsymbol{\beta}$ respectively. We set $\mathrm{vec}(\boldsymbol{\beta}_2) = (\beta_{13}, \beta_{14}, \beta_{15}, \beta_{23}, \beta_{24}, \beta_{25})^{\mathrm{T}} = (0.2, 0.3, 0.4, 0.3, 0.2, 0.4)^{\mathrm{T}}$, and present the estimation results in Table 3 and Figure 3. These results convey the same general message as that of the previous simulation study.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Specifically, when the linearity condition and/or constant variance conditions are satisfied, they are able to contribute to better estimation when the conditions are properly taken into account, in that the estimation efficiency can be further reduced even compared with the optimal estimator without these conditions.

We further experiment with the situation that the covariance of $\mathbf{X}$ is not the identity. We first generate covariates according to models (I), (II) and (III) in the $d = 2$ case. We then multiply the observations by $\mathbf{\Sigma}^{1/2}$ to obtain the correlated $\mathbf{X}$, where $\mathbf{\Sigma}$ is a symmetric matrix, with $0.5^{|i-j|}$ as its $(i, j)$ entry. We present box plots of the squared distances between the estimators and the true parameter values in Figure 4. Because the estimation error of $\mathbf{\Sigma}$ can affect the estimation results significantly when $n = 50$, the differences between the three estimation procedures are not as dramatic as those in the identity covariance case. However, it is clear that Figure 4 still follows the same efficiency improvement patterns as those shown in Figure 3.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Table 1: The absolute biases and standard deviations of the estimators for a logistic $Z$ at $d = 1$. Results based on 1000 simulations at sample size $n = 50$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.

| | (a) | (b) | (c) |
|---|---|---|---|
| | | model (I) | |
| $\beta_2$ | 0.03392 (0.1280) | 0.00614 (0.0694) | 0.02683 (0.2753) |
| $\beta_3$ | 0.06543 (0.1350) | 0.00251 (0.0673) | 0.14387 (0.2850) |
| $\beta_4$ | 0.09790 (0.1416) | 0.01053 (0.0729) | 0.31569 (0.3033) |
| | | model (II) | |
| $\beta_2$ | 0.00252 (0.1071) | 0.00392 (0.1048) | 0.00782 (0.0968) |
| $\beta_3$ | 0.00077 (0.1128) | 0.00342 (0.1094) | 0.00944 (0.0907) |
| $\beta_4$ | 0.00280 (0.1179) | 0.00258 (0.1158) | 0.00389 (0.0996) |
| | | model (III) | |
| $\beta_2$ | 0.04971 (0.1387) | 0.05450 (0.1405) | 0.05322 (0.1897) |
| $\beta_3$ | 0.07423 (0.1395) | 0.08235 (0.1460) | 0.10098 (0.2190) |
| $\beta_4$ | 0.10971 (0.1575) | 0.12779 (0.1673) | 0.16378 (0.2401) |

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Table 2: The absolute biases and standard deviations of the estimators for a gamma $Z$ at $d = 1$. Results based on 1000 simulations at sample size $n = 50$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.

| | (a) | (b) | (c) |
|---|---|---|---|
| | | model (I) | |
| $\beta_2$ | 0.01125 (0.1307) | 0.00126 (0.0474) | 0.25528 (0.5217) |
| $\beta_3$ | 0.02544 (0.1337) | 0.00306 (0.0491) | 0.21482 (0.5554) |
| $\beta_4$ | 0.02010 (0.1263) | 0.00551 (0.0520) | 0.41567 (0.5502) |
| | | model (II) | |
| $\beta_2$ | 0.00383 (0.1011) | 0.00478 (0.0607) | 0.00006 (0.0530) |
| $\beta_3$ | 0.00521 (0.1050) | 0.00476 (0.0607) | 0.00238 (0.0604) |
| $\beta_4$ | 0.01012 (0.1023) | 0.00896 (0.0677) | 0.00630 (0.0644) |
| | | model (III) | |
| $\beta_2$ | 0.01213 (0.1334) | 0.04392 (0.1267) | 0.14229 (0.2748) |
| $\beta_3$ | 0.00381 (0.0998) | 0.03532 (0.0966) | 0.41074 (0.5416) |
| $\beta_4$ | 0.00373 (0.1024) | 0.04272 (0.1067) | 0.44644 (0.6430) |

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

Table 3: The absolute biases and standard deviations of the estimators for $d = 2$. Results based on 1000 simulations at sample size $n = 50$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.

| | (a) | (b) | (c) |
|---|---|---|---|
| | | model (I) | |
| $\beta_{13}$ | 0.01205 (0.1370) | 0.02088 (0.0627) | 0.06876 (0.2062) |
| $\beta_{14}$ | 0.01193 (0.1353) | 0.02780 (0.0733) | 0.03191 (0.2109) |
| $\beta_{15}$ | 0.04068 (0.1511) | 0.03887 (0.0741) | 0.01462 (0.2383) |
| $\beta_{23}$ | 0.08464 (0.3127) | 0.00552 (0.0342) | 0.08232 (0.4441) |
| $\beta_{24}$ | 0.08466 (0.3211) | 0.00874 (0.0346) | 0.08786 (0.4361) |
| $\beta_{25}$ | 0.16100 (0.3478) | 0.01095 (0.0367) | 0.22637 (0.4993) |
| | | model (II) | |
| $\beta_{13}$ | 0.00207 (0.1206) | 0.00205 (0.0670) | 0.00260 (0.0629) |
| $\beta_{14}$ | 0.00216 (0.1127) | 0.00029 (0.0644) | 0.00091 (0.0625) |
| $\beta_{15}$ | 0.00302 (0.1305) | 0.00115 (0.0722) | 0.00363 (0.0646) |
| $\beta_{23}$ | 0.02530 (0.2916) | 0.02511 (0.2240) | 0.00391 (0.1522) |
| $\beta_{24}$ | 0.01126 (0.2636) | 0.00824 (0.2204) | 0.00139 (0.1513) |
| $\beta_{25}$ | 0.04626 (0.2875) | 0.04151 (0.2357) | 0.01934 (0.1679) |
| | | model (III) | |
| $\beta_{13}$ | 0.04646 (0.1294) | 0.08826 (0.1813) | 0.53279 (0.2311) |
| $\beta_{14}$ | 0.02477 (0.1547) | 0.02250 (0.1835) | 0.87108 (0.2697) |
| $\beta_{15}$ | 0.06756 (0.1460) | 0.15399 (0.1892) | 0.49997 (0.3131) |
| $\beta_{23}$ | 0.11495 (0.2415) | 0.36965 (0.3353) | 0.90932 (0.4845) |
| $\beta_{24}$ | 0.12070 (0.2449) | 0.19437 (0.3030) | 0.90814 (0.4303) |
| $\beta_{25}$ | 0.21334 (0.2360) | 0.48398 (0.3812) | 1.63762 (0.5302) |

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

# 7. Discussion

The linearity condition and constant variance condition are common assumptions in the dimension reduction literature. However, their roles and functions became somewhat confusing after a paradox was discovered by Ma and Zhu (2012). Our goal in this article is to provide an intuitive and thorough understanding of these conditions, and hence bring a closure to this mystery. We also provide an optimal way to incorporate these conditions in two new estimators, hence demonstrating the true value of these conditions. The essential messages of this article are the following. First, using the linearity and/or constant variance conditions as a plug-in method as in the classical inverse regression method yields an efficiency loss. Second, this efficiency loss can be decreased by replacing the linearity condition with increasingly more flexible models to capture the mean of the covariates conditional on the reduced dimension covariates, and using OWLS to estimate the parameters of these models. Third, when this series of models becomes maximally flexible, we reach the optimal estimator in this family of estimators, which corresponds to perform dimension reduction without using the linearity or constant variance conditions. Fourth, if we want to take full advantage of the linearity and/or constant variance conditions,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347

plug-in method is not adequate. Analysis under these conditions is needed, and the theoretical results from the new analysis show improved efficiency due to these conditions. Fifth, the newly developed optimal estimators under the linearity and/or constant variance conditions show improvement numerically.

We would like to point out that the linearity puzzle is one of several seemingly puzzling phenomena discovered in statistics; for example, see Henmi et al. (2007); Hitomi et al. (2008) and Kawakita and Kanamori (2013) for several other paradoxes with similar flavor. However, intuitive or geometrical explanations were not given in these works, hence in a sense the understanding of these puzzles is not complete. It will be interesting to investigate the geometrical structure and to further unveil the underneath nature of these puzzles.

# Supplementary Material

The online Supplementary Material includes the conditions required to establish the asymptotic properties of the estimators and detailed proofs of Theorem 1–4.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



## Acknowledgement

The research was supported by US NSF and NIH grants and Hong Kong General Research Fund 27304117.

## References

Amari, S. and M. Kawanabe (1997). Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli 3*, 29–54.

Amari, S. and H. Nagaoka (2007). *Methoos of Information Geometry, Translations of Mathematical Monographs 191.* Oxford: American Mathematical Society/Oxford University Press.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

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

Cook, R. D. and B. Li (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics 30*, 455–474.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



Cook, R. D. and S. Weinberg (1991). Discussion of 'sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association 86*, 28–33.

Henmi, M. and S. Eguchi (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika 91*, 929–941.

Henmi, M., R. Yoshida, and S. Eguchi (2007). Importance sampling via the estimated sampler. *Biometrika 94*, 985–991.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

Hitomi, K., Y. Nishiyama, and R. Okui (2008). A puzzling phenominon in semiparametric estimation problems with infinite-dimentional nuisance parameters. *Econometric Theory 24*, 1717–1728.

Kawakita, M. and T. Kanamori (2013). Semi-supervised learning with density-ratio estimation. *Machine Learning 91*, 189–209.

Li, B. and Y. Dong (2009). Dimension reduction for nonelliptically distributed predictors. *Annals of Statistics 37*, 1272–1298.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



Li, B. and S. Wang (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association 102*, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association 86*, 316–342.

Li, K. C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association 87*, 1025–1039.

Li, K. C. and N. Duan (1989). Regression analysis under link violation. *Annals of Statistics 17*, 1009–1052.

Ma, Y. and L. Zhu (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association 107*, 168–179.

Ma, Y. and L. Zhu (2013a). Efficiency loss caused by linearity condition in dimension reduction. *Biometrika 100*, 371–383.

Ma, Y. and L. Zhu (2013b). Efficient estimation in sufficient dimension reduction. *Annals of Statistics 41*, 250–268.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



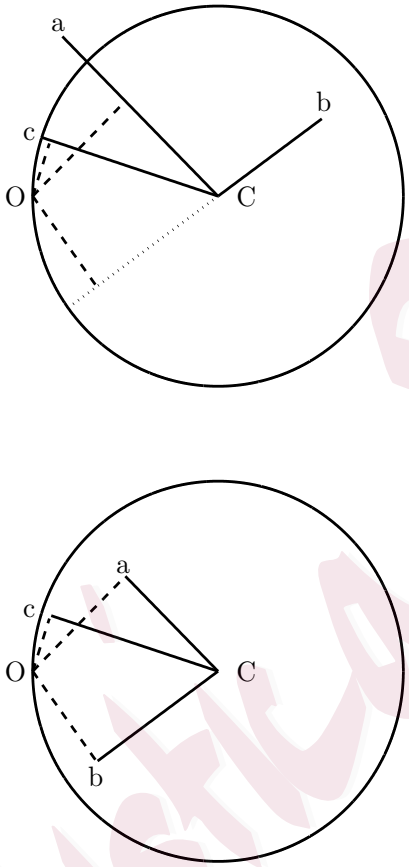


Figure 1: Relative performance of $\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ under WLS (left) and OWLS (right) in terms of their influence functions. $\widehat{\boldsymbol{\beta}}$ is the origin O. $\widetilde{\boldsymbol{\beta}}$ is vector O to circle center C. The vectors from origin O to a, b and c are $\mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$ (left) and $\mathbf{B}_2\mathbf{B}_3^{-1}\mathbf{m}_{\boldsymbol{\alpha}}^{\mathrm{T}}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]$ (right) under three scenarios, resulting in three $\breve{\boldsymbol{\beta}}$ as vectors a to C, b to C and c to C.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



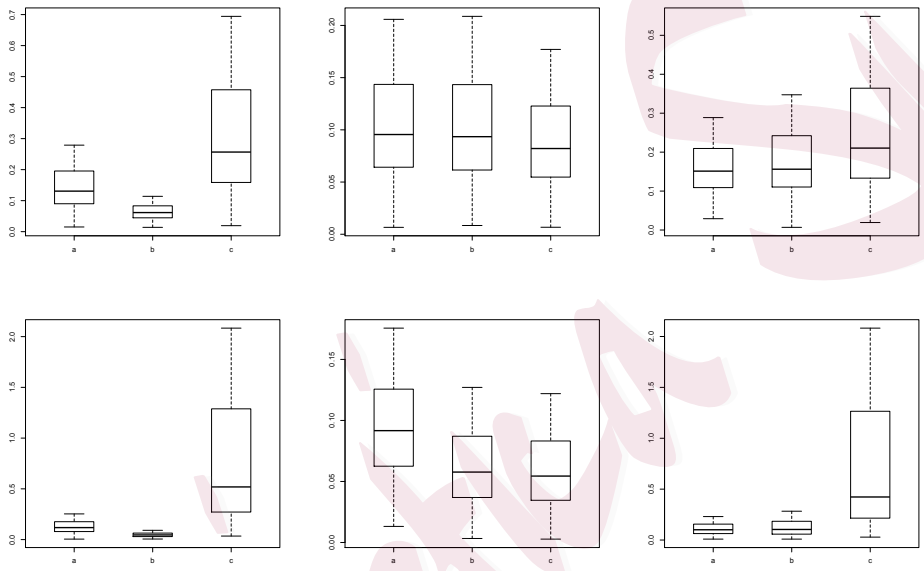

Figure 2: Squared distances of the estimated space to the true space in model (I) (left), model (II) (middle), and model (III) (right) for logistic $Z$ (upper) and gamma $Z$ (lower). The results are based on 1000 simulations with sample size $n = 50$ and $d = 1$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0347



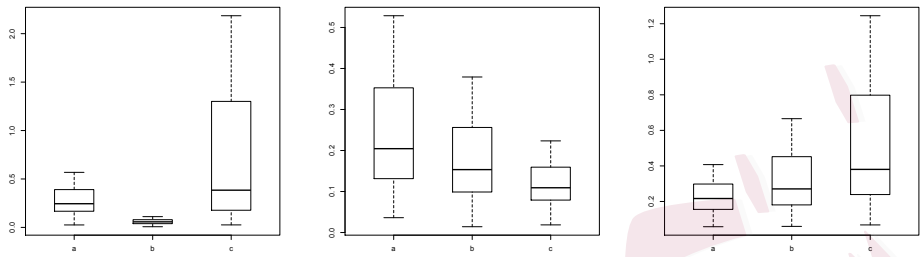

Figure 3: Squared distances of the estimated space to the true space in model (I) (left), model (II) (middle), model (III) (right). Results based on 1000 simulations at sample size $n = 50$, $d = 2$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.

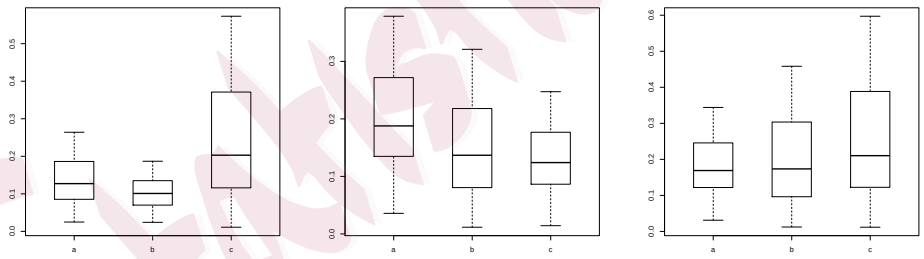

Figure 4: Squared distances of the estimated space to the true space in model (I) (left), model (II) (middle), model (III) (right). Results based 500 simulations of the correlated covariates at sample size $n = 50$ and $d = 2$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator proposed in Theorem 3; (c) is the efficient estimator proposed in Theorem 4.