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: In using inverse regression methods in dimension reduction models, it is discovered that the popular linearity condition has a paradoxical effect: ignoring the linearity condition yields a more efficient estimator than making use of the linearity condition. Through considering classes of parametric models which include the linearity condition as a special case, we examine this phenomenon through a geometry approach and provide an intuitive and extended explanation. Our discovery explains what the real cause of the paradox is, indicates how to properly handle the linearity condition and reveals the true role of the linearity condition. The analysis directly leads to new estimators that further improve the efficient estimator developed earlier in the literature that did not specifically account for the linearity condition and possibly the constant variance condition.

1.   Introduction

Linearity condition, often jointly with constant variance condition, is a popular assumption in 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, i.e. $Y \perp\!\!\!\perp \mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. Linearity condition means the covariate vector $\mathbf{X}$ satisfies that $E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and the constant

---

* co-first

variance condition means that $\text{cov}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ 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 is 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 estimate the relevant quantities, which would be known under these conditions, nonparametrically, then the resulting estimation variance of an inverse regression method will decrease. For brevity, in the following, we call this phenomenon linearity puzzle.

Although the mathematical proof was provided, Ma and Zhu (2013a) did not give any intuition 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 concerning missing values is familiar to many researchers in statistics. There, it was known that in implementing 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 counter-intuitive phenomenon was later beautifully explained via information geometry by Henmi and Eguchi (2004), which reveals the underneath structure of the inverse probability weighting estimator and the inherent reason of the paradox.

This encourages us to also use the tool of information geometry to inspect the linearity puzzle and provide an intuitive explanation in a style similar to that of Henmi and Eguchi (2004). This plan turns out to be partially feasible, in that although we managed to understand and explain the intuition behind the linearity puzzle, the approach we took is a mixture of information geometry and algebra, and is quite different from that of Henmi and Eguchi (2004). Information geometry (Amari and Kawanabe, 1997; Amari and Nagaoka, 2007) usually uses differential geometry tools, here we take it in a wider sense. In the following, we first describe the linearity puzzle in its general form, then inspect a series of models that bridge the completely known function specified in the linearity condition and completely unknown function when this condition is given up in combination with various estimation procedures. Through the investigation of this more general model setting, we are able to imbed the linearity puzzle

2

into a bigger picture and gain the intuitive understanding on why and how it occurs.

We push further the understanding of these phenomena through investigating if 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 via deriving the semiparametric efficiency bound under either or both conditions. Our results indicate that if these conditions are accounted for appropriately, they can indeed contribute to benefitting the estimation in terms of reducing the estimation variability. Encouraged by this discovery, we further devise two new estimators, one is optimal under linearity condition and the other optimal under both linearity and constant variance conditions. We illustrate the gains of the new estimators compared to the efficient estimator proposed in Ma and Zhu (2013b) both theoretically and numerically via a series of simulation studies.

2. Linearity puzzle

Let $\mathbf{X}$ be a $p \times 1$ random vector, Y a univariate random response. In the dimension reduction literature, the original goal of estimating the column space of $\boldsymbol{\beta}$ can be equivalently expressed 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 at identity. A very general class of estimators can be constructed from the estimating function $\mathbf{g}(Y)[\mathbf{a}(\mathbf{X}) - E\{\mathbf{a}(\mathbf{X}) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}}$, $\mathbf{g}$ 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 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 the difference of the asymptotic variances of the two estimators $(\text{cov}(\widetilde{\boldsymbol{\beta}}) - \text{cov}(\widehat{\boldsymbol{\beta}}))$ is positive definite. This result is quite counter intuitive since usually one would expect that additional estimation will inflate the general 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 any parametric models, we consider estimating equations of the form

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

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 the conditional mean of the covariate function $\mathbf{a}(\mathbf{X})$ is plugged in, we name the estimators from (33) as 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 $E(\boldsymbol{\epsilon} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{0}$. Generally speaking, 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 different estimating equations, we use $\widetilde{\boldsymbol{\beta}}$ to denote the estimator from (21),

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

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

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.

<u>Theorem</u> 1. *Under the conditions in the Supplement, the estimators $\widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ satisfy*

$$
\sqrt{n}\boldsymbol{\Sigma}_A \text{vecl}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})
$$

$$
= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]
$$

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

$$
\sqrt{n}\boldsymbol{\Sigma}_A \text{vecl}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})
$$

$$
= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]
$$

$$
+ o_p(1),
$$

$$
\sqrt{n}\boldsymbol{\Sigma}_A \text{vecl}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta})
$$

$$
= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]
$$

5

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a}\otimes E\{\mathbf{g}(Y_i)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\left[\mathbf{a}(\mathbf{X}_i)-\mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i,\boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]$$

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{B}_2\mathbf{B}_1^{-1}\mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\left[\mathbf{a}(\mathbf{X}_i)-\mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i,\boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]+o_p(1).$$

The results regarding $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ were obtained in Ma and Zhu (2013a), while we provide the derivation of the result regarding $\breve{\boldsymbol{\beta}}$ in Supplement S.2. We show the relative performance of the estimators $\widetilde{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$ through their influence functions in Figure 1, left panel.

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 the functions whose difference with $\mathrm{vec}(E\{\mathbf{g}(Y)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X})-E\{\mathbf{a}(\mathbf{X})\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ has the same length as $\mathrm{vec}(E\{\mathbf{g}(Y)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}[\mathbf{a}(\mathbf{X})-E\{\mathbf{a}(\mathbf{X})\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\}]^{\mathrm{T}})$ 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 $\breve{\boldsymbol{\beta}}$, 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 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 and as large as one wishes.

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 $\mathbf{Q}$. However, in general the error variance can depend on $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ as well, so 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 the sense that it has the smallest estimation variance compared to all other WLS estimators. The optimal weight matrix is $\mathbf{Q}^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$. Thus, we now consider exclusively the OWLS estimator $\hat{\boldsymbol{\alpha}}$ 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

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

Define

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

a special case of $\mathbf{B}_1$. The general result concerning $\breve{\boldsymbol{\beta}}$ in Theorem 1 subsequently reduces to a special structure which has a unique orthogonality 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 Supplement S.3.

<u>Theorem 2.</u> *When the OWLS estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta})$ is obtained from solving (45), the resulting estimator $\breve{\boldsymbol{\beta}}$ satisfies*

$$\sqrt{n}\boldsymbol{\Sigma}_A \text{vecl}(\breve{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a} \otimes \left[\mathbf{g}(Y_i) - E\{\mathbf{g}(Y_i) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\right]\left[\mathbf{a}(\mathbf{X}_i) - \mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]$$

7

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{I}_{p_a}\otimes E\{\mathbf{g}(Y_i)\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i\}\left[\mathbf{a}(\mathbf{X}_i)-\mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i,\boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right]$$

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

$$+o_p(1).$$

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 with

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

$$\times\left[\mathbf{a}(\mathbf{X})-\mathbf{m}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X},\boldsymbol{\alpha}_0(\boldsymbol{\beta})\}\right].$$

Similar as in 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

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

is the distance between the center of the circle and an arbitrary line that goes through the origin, hence 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 $\widehat{\boldsymbol{\beta}}$.

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_α is larger, we should use a plane, a hyper-plane, 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 hyper-plane that contains the plane, etc. Correspondingly, the circle will also become a sphere, a hyper-sphere, etc. As a result, the distances between the circle center to the line, plane, hyper-plane 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 nested smaller model" naturally suggests that when the model is "maximized", we should obtain the optimal estimation variance of $\breve{\boldsymbol{\beta}}$. Intuitively, we would think a nonparametric model is the "maximum" parametric model, hence this would explain why the nonparametric based estimator $\widehat{\boldsymbol{\beta}}$ has the smallest variance.

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

9

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 re-
gression kernel estimator.

5. Role and correct use of linearity/constant

variance conditions

Having obtained the results in Section 4 and understood why the nonpara-
metric 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 need to take the approach in Bickel et al. (1993); Tsiatis (2006)
and derive the nuisance tangent space, its orthogonal complement and the
efficient score, for the case with linearity condition and as well as the case
with both linearity and constant variance conditions.

We first investigate the case when only linearity condition is assumed
to hold. We can write 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 $\boldsymbol{\epsilon}_2 = \mathbf{X}_2 - \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \boldsymbol{\beta}_2) = \mathbf{X}_2 - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$,
$\boldsymbol{\beta} = (\mathbf{I}_d, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$. We use f_1, f_2, f_3 to denote the probability density function
(pdf) of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, the pdf of $\boldsymbol{\epsilon}_2$ condition 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 in estimating*

$\boldsymbol{\beta}_2$*, the nuisance tangent space is* $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$*, where*

$$\Lambda_1 = \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} < \infty, \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d} \right],$$

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

$$\Lambda_3 = \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\} < \infty, \right.$$
$$\left. \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d} \right].$$

The nuisance tangent space orthogonal complement is

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

The efficient score function is

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

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 fully take advantage

of this condition and to achieve the 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 in Supplement S.4.

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, 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 a sketch of the proof in Supplement S.5.

<u>Theorem</u> 4. *Assume the linearity and constant variance conditions hold.*

Then in estimating $\boldsymbol{\beta}_2$, the nuisance tangent space is $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$,

where

$$\Lambda_1 = \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})\} < \infty, \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) \in \mathcal{R}^{(p-d)d} \right]$$

$$\Lambda_2 = \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\widetilde{\boldsymbol{\epsilon}}_2 \mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, \right.$$

$$E\{\mathrm{vec}(\widetilde{\boldsymbol{\epsilon}}_2\widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}})\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)\} < \infty,$$

$$\left. \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2) \in \mathcal{R}^{(p-d)d} \right]$$

$$\Lambda_3 = \left[\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) : E\{\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}\} = \mathbf{0}, E\{\mathbf{h}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y)\} < \infty, \right.$$

$$\left. \mathbf{h}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d} \right].$$

5. ROLE AND CORRECT USE OF LINEARITY/CONSTANT VARIANCE CONDITIONS

The nuisance tangent space orthogonal complement is

$$\Lambda^{\perp} \;=\; \Big[\mathbf{g}(\mathbf{X}, Y) : E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2\} = \mathbf{A}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) \widetilde{\boldsymbol{\epsilon}}_2 + \mathbf{B}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) \mathrm{vec}(\widetilde{\boldsymbol{\epsilon}}_2 \widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} - \mathbf{I}_{p-d}),$$

$$E\{\mathbf{g}(\mathbf{X}, Y) \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, Y\} = \mathbf{0}, E\{\mathbf{g}^{\mathrm{T}}(\mathbf{X}, Y) \mathbf{g}(\mathbf{X}, Y)\} < \infty, \mathbf{g}(\mathbf{X}, Y) \in \mathcal{R}^{(p-d)d} \Big].$$

The efficient score function is

$$\mathbf{S}_{\mathrm{eff}}(\mathbf{X}, Y, \boldsymbol{\beta}) \;=\; \mathrm{vec}\left(\mathbf{D}(\boldsymbol{\beta}_2) \widetilde{\boldsymbol{\epsilon}}_2 \frac{\partial \log f_1(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X})}{\partial \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}} + \mathbf{D}(\boldsymbol{\beta}_2) \widetilde{\boldsymbol{\epsilon}}_2 \frac{\partial \log f_3(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, Y)}{\partial \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}} \right)$$

$$-\mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \widetilde{\boldsymbol{\epsilon}}_2 + \mathbf{K}_2(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \mathbf{v} - \mathbf{K}_4(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \mathbf{v}.$$

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

$$\mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \;=\; -\left\{ \frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}} \otimes \mathbf{m}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) + \frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2)}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)} \right\} \{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}}$$

$$\mathbf{K}_2(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \;=\; \mathbf{K}_1(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) E(\mathbf{v}_1 \widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X})^{\mathrm{T}} \left\{ E(\mathbf{v} \mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}) \right\}^{-1}$$

$$\mathbf{C}_1(\boldsymbol{\beta}_2) \;=\; \left[\mathrm{vec}\left\{ \mathbf{D}^{\mathrm{T}}(\boldsymbol{\beta}_2) \frac{\partial \{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}}}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)_1} \right\}, \ldots, \mathrm{vec}\left\{ \mathbf{D}^{\mathrm{T}}(\boldsymbol{\beta}_2) \frac{\partial \{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}}}{\partial \mathrm{vec}(\boldsymbol{\beta}_2)_{(p-d)d}} \right\} \right]^{\mathrm{T}}$$

$$\mathbf{K}_3(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \;=\; \mathbf{C}_1(\boldsymbol{\beta}_2) - \left[\frac{\partial \mathbf{m}^{\mathrm{T}}(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2)}{\partial \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}} \{\mathbf{D}^{-1}(\boldsymbol{\beta}_2)\}^{\mathrm{T}} \right] \otimes \mathbf{D}(\boldsymbol{\beta}_2)$$

$$\mathbf{K}_4(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) \;=\; -\mathbf{K}_3(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \boldsymbol{\beta}_2) E\left[\mathrm{vec}\left\{ \widetilde{\boldsymbol{\epsilon}}_2 \frac{\partial \log f_2(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{X}, \widetilde{\boldsymbol{\epsilon}}_2)}{\partial \widetilde{\boldsymbol{\epsilon}}_2^{\mathrm{T}}} \right\} \mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X} \right] \{E(\mathbf{v} \mathbf{v}^{\mathrm{T}} \mid \boldsymbol{\beta}^{\mathrm{T}} \mathbf{X})\}^{-1}.$$

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 only linearity condition. 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 1 and scale parameter 2. 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 $\mathbf{0}$, 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 0 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 model (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 it 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, while 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), $\mathrm{cov}(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2\{\mathbf{I}_p - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}\}$, which is not a constant matrix, while for model (II), we set $\mathrm{cov}(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = \mathbf{I}_p - \boldsymbol{\beta}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^{\mathrm{T}}$. 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, (c) the efficient estimator proposed in Theorem 4, which assumes both the linearity and the constant variance conditions, and is optimal when

14

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 sample size $n = 50$.

We illustrate the performance of the three estimators in 1000 simulations in Table 1 and 2, where the estimation absolute biases and standard errors are presented. In addition, we create the boxplots for the squared distances of the estimators to the truth, defined by $\{(p - d)d\}^{-1}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ in Figure 2 and 3. From Table 1, 2 and Figure 2, we can see that under model (I) where only the linearity condition is satisfied, estimator (b) has the smaller estimation error compared with estimator (a). This is because estimator (b) utilizes the linearity condition properly to further improve the estimation efficiency from estimator (a). However, estimator (c) goes too far in further assuming the constant variance condition 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 errors. 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 the 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 $\mathbf{u}, \boldsymbol{\epsilon}_2$ in the same way as before. To construct $\mathbf{Z}$, we write the (i, j) element of $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$ as σ_{ij}, and first generate Z_1 from a centered gamma distribution with mean 0 and variance σ_{11}, then 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 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 similar 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) $\tilde{\mathbf{X}}_2 = \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\mathrm{diag}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}) + \{\mathbf{I}_{p-1} - \boldsymbol{\beta}_2(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}_2^{\mathrm{T}}\}^{1/2}\boldsymbol{\epsilon}_2$.

We then let $\mathbf{X}_1 = \mathbf{Z} - \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{X}_2$, $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$ in models (I) and (II). For model (III), we let $\tilde{\mathbf{X}}_1 = \mathbf{Z} - \boldsymbol{\beta}_2^{\mathrm{T}}\tilde{\mathbf{X}}_2$, $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1^{\mathrm{T}}, \tilde{\mathbf{X}}_2^{\mathrm{T}})^{\mathrm{T}}$ and form $\mathbf{X} = \{\mathrm{var}(\tilde{\mathbf{X}})\}^{-1/2}\tilde{\mathbf{X}}$. Some careful algebraic calculation can then be employed to verify that the construction in the three models satisfies $E(\mathbf{X}) = \mathbf{0}$, $\mathrm{var}(\mathbf{X}) = \mathbf{I}_p$. In addition, $\mathbf{X}$ satisfies only the linearity condition in model (I), it further satisfies the constant variance condition in model (II), and it does not

15

satisfy the linearity condition in model (III). We then generate Y from a normal distribution with mean $\sin(2\boldsymbol{\beta}_{.1}^{\mathrm{T}}\mathbf{X})$ and variance $\log\{2 + (\boldsymbol{\beta}_{.2}^{\mathrm{T}}\mathbf{X})^2\}$, where $\boldsymbol{\beta}_{.1}, \boldsymbol{\beta}_{.2}$ are respectively the first and second columns of $\boldsymbol{\beta}$. 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 in the previous simulation study. Specifically, when 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 to the optimal estimator without these conditions.

We further experiment with the situation that the covariance of $\mathbf{X}$ is not identity. We first generate covariates according to model (I), (II), (III) in the $d = 2$ case. We then multiply the observations by $\boldsymbol{\Sigma}^{1/2}$ to obtain the correlated $\mathbf{X}$, where $\boldsymbol{\Sigma}$ is a symmetric matrix with $0.5^{|i-j|}$ as its (i, j) entry. We present the boxplots of the squared distances between the estimators and the true parameter values in Figure 4. Because the estimation error of $\boldsymbol{\Sigma}$ can considerably affect the estimation results when $n = 50$, the differences among 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.

7. Discussion

Linearity condition and constant variance condition are common assumptions in dimension reduction literature. Their roles and functions become somewhat confusing after the paradox discovered in 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 demonstrate the real value of these conditions. The essential messages of this article are the following. First, linearity and/or constant variance conditions used as a plug-in method as in the classical inverse regression method brings an efficiency loss. Second, this efficiency loss can be decreased via replacing the linearity condition by increasingly more flexible models to capture the mean of the covariates conditional on the reduced dimension covariates, and using optimal weighted least square to estimate

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)		
β_2	0.03392 (0.1280)	0.00614 (0.0694)	0.02683 (0.2753)
β_3	0.06543 (0.1350)	0.00251 (0.0673)	0.14387 (0.2850)
β_4	0.09790 (0.1416)	0.01053 (0.0729)	0.31569 (0.3033)
	model (II)		
β_2	0.00252 (0.1071)	0.00392 (0.1048)	0.00782 (0.0968)
β_3	0.00077 (0.1128)	0.00342 (0.1094)	0.00944 (0.0907)
β_4	0.00280 (0.1179)	0.00258 (0.1158)	0.00389 (0.0996)
	model (III)		
β_2	0.04971 (0.1387)	0.05450 (0.1405)	0.05322 (0.1897)
β_3	0.07423 (0.1395)	0.08235 (0.1460)	0.10098 (0.2190)
β_4	0.10971 (0.1575)	0.12779 (0.1673)	0.16378 (0.2401)

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)		
β_2	0.01125 (0.1307)	0.00126 (0.0474)	0.25528 (0.5217)
β_3	0.02544 (0.1337)	0.00306 (0.0491)	0.21482 (0.5554)
β_4	0.02010 (0.1263)	0.00551 (0.0520)	0.41567 (0.5502)
	model (II)		
β_2	0.00383 (0.1011)	0.00478 (0.0607)	0.00006 (0.0530)
β_3	0.00521 (0.1050)	0.00476 (0.0607)	0.00238 (0.0604)
β_4	0.01012 (0.1023)	0.00896 (0.0677)	0.00630 (0.0644)
	model (III)		
β_2	0.01213 (0.1334)	0.04392 (0.1267)	0.14229 (0.2748)
β_3	0.00381 (0.0998)	0.03532 (0.0966)	0.41074 (0.5416)
β_4	0.00373 (0.1024)	0.04272 (0.1067)	0.44644 (0.6430)

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

the parameters of these models. Third, when this series of models become 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 fully take advantage of the linearity and/or constant variance conditions, 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.

8. Supplementary material

The supplementary material includes the conditions to establish the asymptotic properties of the estimators and the detailed proofs of Theorem 1–4.

9. Acknowledgement

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.

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.

21

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.

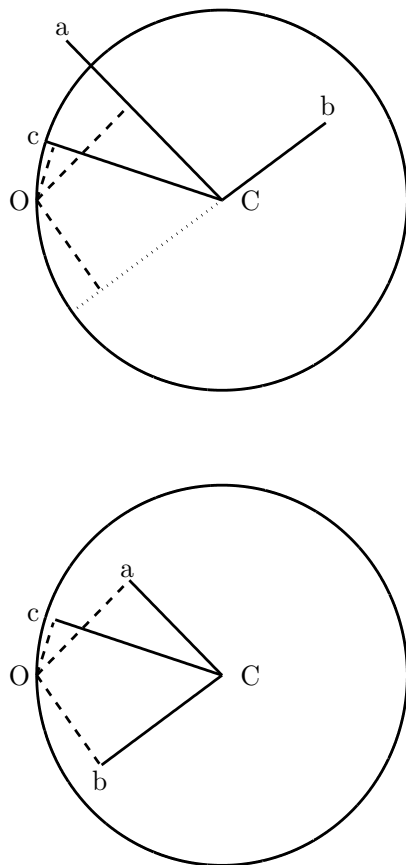Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Figure 1: Relative performance of $\widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}$ and $\widecheck{\boldsymbol{\beta}}$ under WLS (left) and OWLS (right) in terms of their influence functions. $\widehat{\boldsymbol{\beta}}$ is 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 $\widecheck{\boldsymbol{\beta}}$ as vectors a to C, b to C and c to C.
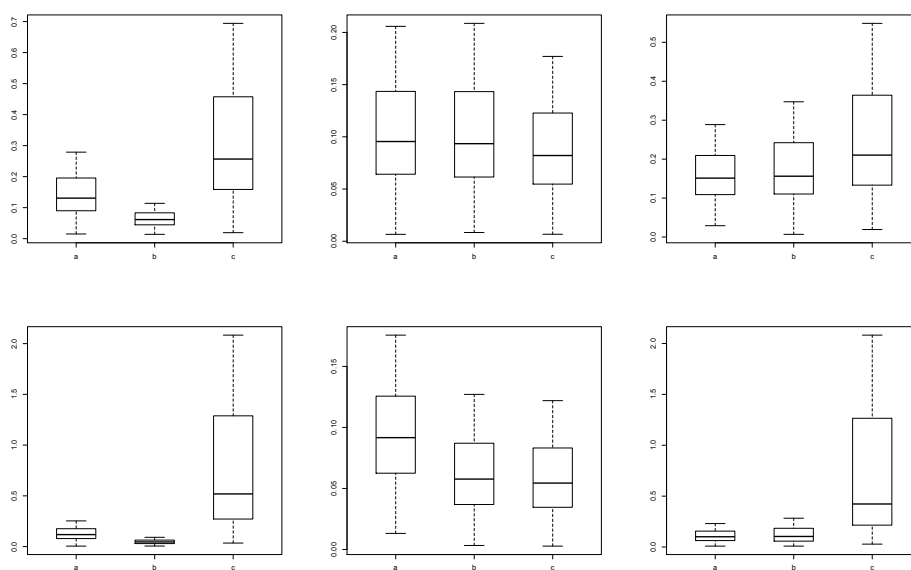
23

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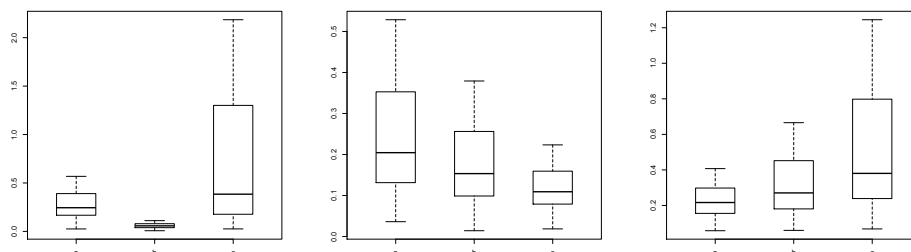
Figure 3: Squared distances of the estimated space to the true space in model (I) (left), model (II) (middle), model (III) (right). Results based on 1000 simulations at sample size $n = 50$, $d = 2$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is the efficient estimator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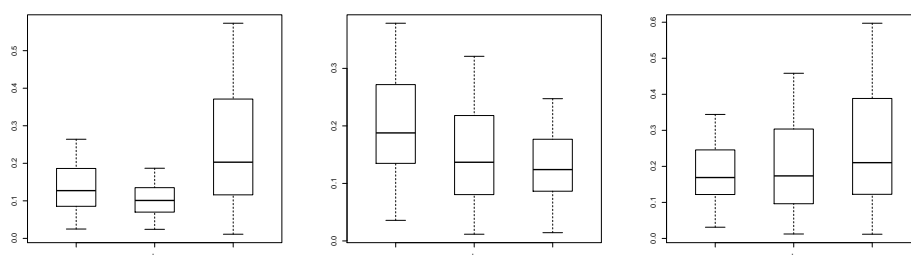

Figure 4: Squared distances of the estimated space to the true space in model (I) (left), model (II) (middle), model (III) (right). Results based 500 simulations of the correlated covariates at sample size $n = 50$, $d = 2$. Estimator (a) is the efficient estimator given in Ma and Zhu (2013a); (b) is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