# Conditional Variance Estimator for Sufficient Dimension Reduction

**Lukas Fertl**[*]
Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
TU Wien, Vienna, Austria

**Efstathia Bura** [†]
Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
TU Wien, Vienna, Austria

February 18, 2021

**ABSTRACT**

*Conditional Variance Estimation* (CVE) is a novel sufficient dimension reduction (SDR) method for additive error regressions with continuous predictors and link function. It operates under the assumption that the predictors can be replaced by a lower dimensional projection without loss of information. In contrast to the majority of moment based sufficient dimension reduction methods, Conditional Variance Estimation is fully data driven, does not require the restrictive linearity and constant variance conditions, and is not based on inverse regression. CVE is shown to be consistent and its objective function to be uniformly convergent. CVE outperforms the mean average variance estimation, (MAVE), its main competitor, in several simulation settings, remains on par under others, while it always outperforms the usual inverse regression based linear SDR methods, such as Sliced Inverse Regression.

## 1 Introduction

Suppose $(Y, \mathbf{X}^T)^T$ have a joint continuous distribution, where $Y \in \mathbb{R}$ denotes a univariate response and $\mathbf{X} \in \mathbb{R}^p$ a $p$-dimensional covariate vector. We assume that the dependence of $Y$ and $\mathbf{X}$ is modelled by

$$Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon, \tag{1}$$

where $\mathbf{X}$ is independent of $\epsilon$ with positive definite variance-covariance matrix, $\mathbb{V}\mathrm{ar}(X) = \mathbf{\Sigma_x}$, $\epsilon \in \mathbb{R}$ is a mean zero random variable with finite $\mathbb{V}\mathrm{ar}(\epsilon) = E\left(\epsilon^2\right) = \eta^2$, $g$ is an unknown continuous non-constant function, and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathbb{R}^{p\times k}$ of rank $k \leq p$. Model (1) states that

$$\mathbb{E}(Y \mid \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{X}) \tag{2}$$

and requires the first conditional moment $\mathbb{E}(Y \mid \mathbf{X}) = g(\mathbf{B}^T\mathbf{X})$ contain the entirety of the information in $X$ about $Y$ and be captured by $\mathbf{B}^T\mathbf{X}$, so that $F(Y \mid \mathbf{X}) = F(Y \mid \mathbf{B}^T\mathbf{X})$, where $F(\cdot \mid \cdot)$ denotes the conditional cumulative distribution function (cdf) of the first given the second argument. That is, $Y$ is statistically independent of $\mathbf{X}$ when $\mathbf{B}^T\mathbf{X}$ is given and replacing $\mathbf{X}$ by $\mathbf{B}^T\mathbf{X}$ induces no loss of information for the regression of $Y$ on $\mathbf{X}$.

[*]lukas.fertl@tuwien.ac.at
[†]efstathia.bura@tuwien.ac.at

Identifying the span of $\mathbf{B}$; i.e., the column space of $\mathbf{B}$, as only the $\text{span}\{\mathbf{B}\}$ is identifiable, suffices in order to identify the *sufficient reduction* of $\mathbf{X}$ for the regression of $Y$ on $\mathbf{X}$. We assume, without loss of generality, $\mathbf{B}$ is semi-orthogonal, i.e., $\mathbf{B}^T\mathbf{B} = \mathbf{I}_k$, since a change of coordinate system by an orthogonal transformation does not alter model (2).

For $q \leq p$, let

$$\mathcal{S}(p,q) = \{\mathbf{V} \in \mathbb{R}^{p\times q} : \mathbf{V}^T\mathbf{V} = \mathbf{I}_q\}, \tag{3}$$

denote the Stiefel manifold, that comprizes of all $p \times q$ matrices with orthonormal columns. $\mathcal{S}(p,q)$ is compact and $\dim(\mathcal{S}(p,q)) = pq - q(q+1)/2$ [see [4] and Section 2.1 of [31]]. Further let

$$Gr(p,q) = \mathcal{S}(p,q)/\mathcal{S}(q,q) \tag{4}$$

denote the Grassmann manifold, i.e. all $q$-dimensional subspaces in $\mathbb{R}^p$, which is exactly the quotient space of $\mathcal{S}(p,q)$ with all $q \times q$ orthonormal matrices $\mathcal{S}(q,q)$, i.e. the basis of a linear subspace is unique up to orthogonal transformations.

The fact that only $\text{span}\{\mathbf{B}\}$ is identifiable, can be expressed through the Grassmann manifold $Gr(p,q)$ in (4). The goal of sufficient dimension reduction in model (1) is to find a subspace $\mathbf{M} \in Gr(p,k)$ such that any basis $\mathbf{B} \in \mathcal{S}(p,k)$ of $\mathbf{M}$ fulfills (1) or equivalently (2).

Finding sufficient reductions of the predictors to replace them in regression and classification without loss of information is called *sufficient dimension reduction* [9]. The first split in sufficient dimension reduction taxonomy occurs between likelihood and non-likelihood based methods. The former, which were developed more recently [11, 10, 12, 6, 5], assume knowledge either of the joint family of distributions of $(Y, \mathbf{X}^T)^T$, or the conditional family of distributions for $\mathbf{X} \mid Y$. The latter is the most researched branch of sufficient dimension reduction and comprizes of three classes of methods: Inverse regression based, semi-parametric and nonparametric. Reviews of the former two classes can be found in [1, 25, 22].

In this paper we present the *conditional variance estimation*, which falls in the class of nonparametric methods. The estimators in this class minimize a criterion that describes the fit of the dimension reduction model (2) under (1) to the observed data. Since the criterion involves unknown distributions or regression functions, nonparametric estimation is used to recover $\text{span}\{\mathbf{B}\}$. Statistical approaches to identify $\mathbf{B}$ in (2) include ordinary least squares and nonparametric multiple index models [34]. The least squares estimator, $\mathbf{\Sigma}_{\mathbf{x}}^{-1}\text{cov}(\mathbf{X}, Y)$, always falls in $\text{span}\{\mathbf{B}\}$ [22, Th. 8.3]. Principal Hessian Directions [24] was the first sufficient dimension reduction estimator to target $\text{span}\{\mathbf{B}\}$ in (2). Its main disadvantage is that it requires the so called *linearity* and *constant variance* conditions on the marginal distribution of $\mathbf{X}$. Its relaxation, Iterative Hessian Transformation [13], still requires the linearity condition in order to recover vectors in $\text{span}\{\mathbf{B}\}$.

The most competitive nonparametric sufficient dimension reduction method up to now has been *minimum average variance estimation* (MAVE, [35]). It assumes model (1), bounded fourth derivative covariate density, and existence of continuous bounded third derivatives for $g$. It uses a local first order approximation of $g$ in (1) and minimizes the expected conditional variance of the response given $\mathbf{B}^T\mathbf{X}$.

The *conditional variance estimator* also targets and recovers $\text{span}\{B\}$ in models (1) and (2). The objective function is based on the intuition that the directions in the predictor space that capture the dependence of $Y$ on $X$ should exhibit significantly higher variation in $Y$ as compared with the directions along which $Y$ exhibits markedly less variation. The *conditional variance estimator* is a fully data-driven estimator that performs better than or is on par with *minimum average variance estimation* in simulations. The *conditional variance estimator* differs from other approaches, including MAVE, in that it only targets the $\text{span}\{\mathbf{B}\}$ and does not require an explicit form or estimation of the link function $g$. As a result, it requires weaker assumptions on its smoothness.

## 2 Motivation

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\mathbf{X} : \Omega \to \mathbb{R}^p$ be a random vector with a continuous probability density function $f_{\mathbf{X}}$ and denote its support by $\text{supp}(f_{\mathbf{X}})$. Throughout $\|\cdot\|$ denotes the Frobenius norm for matrices, Euclidean norm for vectors, and scalar product refers to the euclidean scalar product. For any matrix $\mathbf{M}$, or linear subspace $\mathbf{M}$, we denote by $\mathbf{P}_{\mathbf{M}}$ the projection matrix on the column space of the matrix or on the subspace, i.e. $\mathbf{P}_{\mathbf{M}} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \in \mathbb{R}^{p\times p}$ for $\mathbf{M} \in \mathbb{R}^{p\times q}$. For any $\mathbf{V} \in \mathcal{S}(p,q)$, defined in (3), we generically denote a basis of the orthogonal complement of its column space $\text{span}\{\mathbf{V}\}$, by $\mathbf{U}$. That is, $\mathbf{U} \in \mathcal{S}(p, p-q)$ such that $\text{span}\{\mathbf{V}\} \perp \text{span}\{U\}$ and $\text{span}\{\mathbf{V}\} \cup \text{span}\{\mathbf{U}\} = \mathbb{R}^p$, $\mathbf{U}^T\mathbf{V} = \mathbf{0} \in \mathbb{R}^{(p-q)\times q}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{p-q}$. For any $\mathbf{x}, \mathbf{s}_0 \in \mathbb{R}^p$ we can always write

$$\mathbf{x} = \mathbf{s}_0 + \mathbf{P}_{\mathbf{V}}(\mathbf{x} - \mathbf{s}_0) + \mathbf{P}_U(\mathbf{x} - \mathbf{s}_0) = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2 \tag{5}$$

where $\mathbf{r}_1 = \mathbf{V}^T(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^q$, $\mathbf{r}_2 = \mathbf{U}^T(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^{p-q}$.

In the sequel, we refer to the following assumptions as needed and the proofs of the Theorems are presented in the Appendix.

**(A.1).** *Model* $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$ *holds with* $Y \in \mathbb{R}$, $g : \mathbb{R}^k \to \mathbb{R}$ *non constant in all arguments,* $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathbb{R}^{p\times k}$ *of rank* $k \leq p$, $\mathbf{X} \in \mathbb{R}^p$ *independent from* $\epsilon$, $\mathbb{V}\mathrm{ar}(\mathbf{X}) = \mathbf{\Sigma}_\mathbf{x}$ *is positive definite ,* $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}\mathrm{ar}(\epsilon) = \eta^2 < \infty$.

**(A.2).** *The link function* $g$ *and the density* $f_\mathbf{X} : \mathbb{R}^p \to [0, \infty)$ *of* $\mathbf{X}$ *are twice continuous differentiable.*

**(A.3).** $\mathbb{E}(|Y|^8) < \infty$.

**(A.4).** *supp*$(f_\mathbf{X})$ *is compact.*

**Remark.** *Assumption (A.4) is not as restrictive as it might seem. [36] showed in Proposition 11 that there is a compact set* $\mathcal{S} \subset \mathbb{R}^p$ *such that the mean subspace of model* (1) *is the same as the mean subspace of* $Y = g(\mathbf{B}^T\mathbf{X}_{|\mathcal{S}}) + \epsilon$, *where* $\mathbf{X}_{|\mathcal{S}} = \mathbf{X}1_{\{\mathbf{X}\in\mathcal{S}\}}$ *and* $1_A$ *is the indicator function of* $A$. *Further* $\mathcal{S}$ *can be assumed to be an ellipsoid and for all* $\widetilde{\mathcal{S}} \supseteq \mathcal{S}$ *the same assertion holds true.*

**Definition.** *For* $q \leq p \in N$ *and any* $\mathbf{V} \in \mathcal{S}(p, q)$, *we define*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mathbb{V}\mathrm{ar}(Y \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}), \tag{6}$$

*where* $\mathbf{s}_0 \in \mathbb{R}^p$ *is a shifting point.*

**Definition.** *For* $\mathbf{V} \in \mathcal{S}(p, q)$, *we define the objective function,*

$$L(\mathbf{V}) = \int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}) f_\mathbf{X}(\mathbf{x}) d\mathbf{x} = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right). \tag{7}$$

$L(\mathbf{V})$ in (7) is the objective function for the estimator we propose for the span of $\mathbf{B}$ in (1) and Theorem 1 provides the statistical motivation for the objective function (7) of the conditional variance estimator. First we derive that both population based functions (6) and (7) are well defined.

Let $\mathbf{X}$ be a $p$-dimensional continuous random vector with density $f_\mathbf{X}(\mathbf{x})$, $\mathbf{s}_0 \in \mathrm{supp}(f_\mathbf{X}) \subset \mathbb{R}^p$, and $\mathbf{V}$ belongs to the Stiefel manifold $\mathcal{S}(p, q)$ defined in (3). The function

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\mathrm{span}\{\mathbf{V}\}}(\mathbf{r}_1) = \frac{f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} \tag{8}$$

is a proper conditional density of $\mathbf{X}$ that is concentrated on the affine subspace $\mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}$ using the concept of regular conditional probability [21] under assumption (A.2). The detailed justification is given in the Appendix, where we also show that under assumptions (A.1), (A.2) and (A.4), $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (6) and $L(\mathbf{V})$ in (7) are well defined and continuous. Moreover,

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 + \eta^2 \tag{9}$$

where

$$\mu_l(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1)^l \frac{f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} d\mathbf{r}_1 = \frac{t^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0)} \tag{10}$$

with

$$t^{(l)}(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1)^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1. \tag{11}$$

**Theorem 1.** *Suppose* $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q) \in \mathcal{S}(p, q)$ *and* $q \in \{1, \ldots, p\}$. *Under assumptions (A.1), (A.2) and (A.4),*

*(a) For all* $\mathbf{s}_0 \in \mathbb{R}^p$ *and* $\mathbf{V}$ *such that there exist* $u \in \{1, ..., q\}$ *with* $\mathbf{v}_u \in \mathrm{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \mathbb{V}\mathrm{ar}(\epsilon) = \eta^2$ *and* $L(\mathbf{V}) > \eta^2$.

*(b) For all* $\mathbf{s}_0 \in \mathbb{R}^p$ *and* $\mathrm{span}\{\mathbf{V}\} \perp \mathrm{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \eta^2$ *and* $L(\mathbf{V}) = \eta^2$.

*Proof.* Let $\mathbf{s}_0 \in \mathbb{R}^p$ and $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q) \in \mathbb{R}^{p\times q}$ so that $\mathbf{v}_u \in \mathrm{span}\{\mathbf{B}\}$ for some $u \in \{1, ..., q\}$. To obtain (a), observe $\mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\} \iff \mathbf{X} = \mathbf{s}_0 + \mathbf{P}_V(\mathbf{X} - \mathbf{s}_0)$ and using (6) yields

$$\begin{aligned}\tilde{L}(\mathbf{V}, \mathbf{s}_0) &= \mathbb{V}\mathrm{ar}\left(g(\mathbf{B}^T\mathbf{X}) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)\right) + \mathbb{V}\mathrm{ar}(\epsilon) \\ &= \mathbb{V}\mathrm{ar}\left(g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)\right) + \eta^2 > \eta^2 \end{aligned} \tag{12}$$

since $\mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \neq 0$ with probability 1, and therefore the variance term in (12) is positive. For $\mathbf{V}$ such that $\mathbf{V}$ and $\mathbf{B}$ are orthogonal, $\mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) = 0$ and (b) follows. Since $\mathbf{s}_0$ is arbitrary yet constant, the statements for $L(\mathbf{V})$ follow. □

Theorem 1 also has an intuitive geometrical interpretation for the proposed method. If $\mathbf{X}$ is not random, the deterministic function $Y = g(\mathbf{B}^T\mathbf{X})$ is constant in all directions orthogonal to $\mathbf{B}$ and varies in all other directions. If randomness is introduced, as in model (1), then the variation in $Y$ stems only from $\epsilon$ in all directions orthogonal to $\mathbf{B}$. In all other directions the variation comprizes of the sum of the variation of $\epsilon$ and of $g(\mathbf{B}^T\mathbf{X})$. In consequence, the objective function (7) captures the variation of $Y$ as $\mathbf{X}$ varies in the column space of $\mathbf{V}$ and is minimized in the directions orthogonal to $\mathbf{B}$.

### 2.1 Conditional Variance Estimator (CVE)

We have shown that the objective function $L(\mathbf{V})$ in (7) is well defined and continuous in Section 2. Let

$$\mathbf{V}_q = \operatorname{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L(\mathbf{V}). \tag{13}$$

$\mathbf{V}_q$ is well defined as the minimizer of a continuous function over the compact set $\mathcal{S}(p,q)$. Nevertheless, $\mathbf{V}_q$ is not unique since for all orthogonal $\mathbf{O} \in \mathbb{R}^{q\times q}$ such that $\mathbf{O}\mathbf{O}^T = \mathbf{I}_q$, $L(\mathbf{V}\mathbf{O}) = L(\mathbf{V})$ as $L(\mathbf{V})$ depends on $\mathbf{V}$ only through $\operatorname{span}\{\mathbf{V}\}$. Nevertheless, it is a unique minimizer over the Grassmann manifold $Gr(p,q)$ in (4). To see this, suppose $\mathbf{V} \in \mathcal{S}(p,q)$ is an arbitrary basis of a subspace $\mathbf{M} \in Gr(p,q)$. We can identify $M$ through the projection $\mathbf{P}_\mathbf{M} = \mathbf{V}\mathbf{V}^T$. By (5) we write $\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$. By the Fubini-Tornelli Theorem we obtain

$$\begin{aligned}\tilde{t}^{(l)}(\mathbf{P}_\mathbf{M}, \mathbf{s}_0) &= \int_{\text{supp}(f_\mathbf{X})} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T P_\mathbf{M}\mathbf{x})^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{P}_\mathbf{M}\mathbf{x}) d\mathbf{x} \\ &= t^{(l)}(\mathbf{V}, \mathbf{s}_0) \int_{\text{supp}(f_\mathbf{X})\cap\mathbb{R}^{p-q}} d\mathbf{r}_2.\end{aligned} \tag{14}$$

Therefore $\tilde{t}^{(l)}(\mathbf{P}_\mathbf{M}, \mathbf{s}_0)/\tilde{t}^{(0)}(\mathbf{P}_\mathbf{M}, \mathbf{s}_0) = t^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ and $\mu_l(\cdot, \mathbf{s}_0)$ in (10) can also be viewed as a function from $Gr(p,q)$ to $\mathbb{R}$. If the optimization (13) is over $Gr(p,q)$, the objective function (7) has a unique minimum at $\operatorname{span}\{\mathbf{B}\}^\perp$ by Theorem 1. Therefore $\mathbf{B}$ is not uniquely identifiable but its $\operatorname{span}\{\mathbf{B}\}$ is.

Corollary 2 follows directly from Theorem 1 and provides the means for identifying the linear projections of the predictors satisfying (1).

**Corollary 2.** *Under the assumptions (A.1), (A.2), and (A.3) the solution of the optimisation problem* $\mathbf{V}_q$ *in* (13) *is well defined. Let* $k = \dim(\operatorname{span}\{\mathbf{B}\})$ *and* $q = p - k$*,*

*(a)* $\operatorname{span}\{\mathbf{V}_q\} = \operatorname{span}\{\mathbf{B}\}^\perp$

*(b)* $\operatorname{span}\{\mathbf{V}_q\}^\perp = \operatorname{span}\{\mathbf{B}\}$

We next define the novel estimator of the sufficient reduction space, $\operatorname{span}\{\mathbf{B}\}$, in (1), which is motivated by Theorem 1 and Corollary 2 (b) serves as the estimation equation for the conditional variance estimator at the population level.

**Definition.** *The* ***Conditional Variance Estimator*** *is defined to be any basis* $\mathbf{B}_{p-q}$ *of* $\operatorname{span}\{\mathbf{V}_q\}^\perp$*. That is, the CVE of* $\mathbf{B}$ *is any* $\mathbf{B}_{p-q}$ *such that*

$$\operatorname{span}\{\mathbf{B}_{p-q}\} = \operatorname{span}\{\mathbf{V}_q\}^\perp \tag{15}$$

When $q = p - k$, where $k = \text{rank}(\mathbf{B})$ in (1), then the CVE obtains the population $\operatorname{span}\{\mathbf{B}\}$. Alternatively, we can also target $\mathbf{B}$ directly by maximizing the objective function $L(\mathbf{V})$. The downside of this approach is that $\mathbf{X}$ either needs to be standardized, or the conditioning argument needs to be changed to $\mathbf{X} = \mathbf{s}_0 + \mathbf{P}_{\mathbf{\Sigma}_\mathbf{x}^{-1}(\operatorname{span}\{\mathbf{V}\})}(\mathbf{X} - \mathbf{s}_0)$, where $\mathbf{P}_{\mathbf{M}(\operatorname{span}\{\mathbf{V}\})}$ is the orthogonal projection operator with respect to the inner product $\langle \mathbf{x}, \mathbf{y}\rangle_\mathbf{M} = \mathbf{x}^T\mathbf{M}\mathbf{y}$. In either case, the inversion of $\mathbf{\Sigma}_\mathbf{x}$ is required. Our choice of targeting the orthogonal complement avoids the inversion of $\mathbf{\Sigma}_\mathbf{x}$, and the estimation algorithm in Section 4 can be applied to regressions with $p > n$ or $p \approx n$, where $n$ denotes the sample size. Additionally, targeting the complement has computational advantages. The dimension of the search space $\operatorname{span}\{\mathbf{V}_q\}^\perp$ is $p - q$, which is smaller than the dimension of the direct target space in (15) when $q = p - k$ for small $k$, which is the appropriate setting in a dimension reduction context.

## 3 Estimation

Assume $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}^T$ is an independent identical distributed sample from model (1). For $\mathbf{V} \in \mathcal{S}(p,q)$ and $\mathbf{s}_0 \in \mathbb{R}^p$, we define

$$\begin{aligned}d_i(\mathbf{V}, \mathbf{s}_0) &= \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}}\mathbf{X}_i\|^2 = \|\mathbf{X}_i - \mathbf{s}_0\|^2 - \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{V}\mathbf{V}^T(\mathbf{X}_i - \mathbf{s}_0)\rangle \\ &= \|(\mathbf{I}_p - \mathbf{V}\mathbf{V}^T)(\mathbf{X}_i - \mathbf{s}_0)\|^2 = \|\mathbf{P}_\mathbf{U}(\mathbf{X}_i - \mathbf{s}_0)\|^2\end{aligned} \tag{16}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P_V} = \mathbf{V}\mathbf{V}^T$ and $\mathbf{P}_U = \mathbf{I}_p - \mathbf{P_V}$ using the orthogonal decomposition given by (5).

Let $h_n \in \mathbb{R}_+$ be a sequence of bandwidths and we call the set $\mathcal{S}_{\mathbf{s}_0,\mathbf{V}} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$ a *slice* that depends on both the shifting point $\mathbf{s}_0$ and the matrix $\mathbf{V}$. $h_n$ represent the squared width of a slice around the subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ and fulfills the following assumptions.

**(H.1).** *For $n \to \infty$, $h_n \to 0$*

**(H.2).** *For $n \to \infty$, $nh_n^{(p-q)/2} \to \infty$*

**Remark.** *For obtaining the consistency of the proposed estimator (H.2) will be strengthened to* $\log(n)/nh_n^{(p-q)/2} \to 0$.

Let $K$ be a function satisfying the following assumptions.

**(K.1).** *$K : [0, \infty) \to [0, \infty)$ is a non increasing and continuous function, so that $|K(z)| \leq M_1$, with $\int_{\mathbb{R}^q} K(\|\mathbf{r}\|^2)d\mathbf{r} < \infty$ for $q \leq p - 1$.*

**(K.2).** *There exist positive finite constants $L_1$ and $L_2$ such that the kernel $K$ satisfies one of the following:*

*(1) $K(u) = 0$ for $|u| > L_2$ and for all $u, \tilde{u}$ it holds $|K(u) - K(\tilde{u})| \leq L_1|u - \tilde{u}|$*

*(2) $K(u)$ is differentiable with $|\partial_u K(u)| \leq L_1$ and for some $\nu > 1$ it holds $|\partial_u K(u)| \leq L_1|u|^{-\nu}$ for $|u| > L_2$*

Examples of functions that satisfy (K.1) and (K.2) include the Gaussian, $K(z) = c\exp(-z^2/2)$, the exponential, $K(z) = c\exp(-z)$, and the squared Epanechnikov kernel, $K(z) = c\max\{(1 - z^2), 0\}^2$ (i.e. polynomial kernels), where $c$ is a constant. The rectangular, $K(z) = cI(z \leq 1)$, does not fulfill the assumptions but will be mentioned for intuitive explanations. A list of further kernel functions is given in [28, Table 1].

### 3.1 The estimator of $L(\mathbf{V})$ and its uniform convergence

**Definition.** *For $i = 1, \ldots, n$, we define*

$$w_i(\mathbf{V}, \mathbf{s}_0) = \frac{K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{d_j(\mathbf{V},\mathbf{s}_0)}{h_n}\right)} \tag{17}$$

**Definition.** *The sample based estimate of $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ is defined as*

$$\tilde{L}_n(\mathbf{V}, s_0) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0)(Y_i - \bar{y}_1(\mathbf{V}, \mathbf{s}_0))^2 = \bar{y}_2(\mathbf{V}, \mathbf{s}_0) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0)^2 \tag{18}$$

*where $\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0)Y_i^l$, $l = 1, 2$.*

**Definition.** *The estimate of the objective function $L(\mathbf{V})$ in (7) is defined as*

$$L_n(\mathbf{V}) = \frac{1}{n}\sum_{i=1}^n \tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \tag{19}$$

*where each data point $\mathbf{X}_i$ is a shifting point.*

To obtain insight as to the choice of $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ in (18), let us consider the rectangular kernel, $K(z) = 1_{\{z \leq 1\}}$. In this case, $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ computes the empirical variance of the $Y_i$'s corresponding to the $\mathbf{X}_i$'s that are no further than $\sqrt{h_n}$ away from the affine space $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, i.e., $d_i(\mathbf{V}, \mathbf{s}_0) = \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{X}_i\|^2 \leq h_n$. If a smooth kernel is used, such as the Gaussian in our simulation studies, then $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ is also smooth, which allows the computation of gradients required to solve the optimization problem.

In Theorem 3 we state the conditions under which $L_n(\mathbf{V})$ in (19) converges uniformly to its population counterpart in (7). This result will lead to the consistency of our estimator.

**Theorem 3.** *Let $\tilde{a}_n^2 = \log(n)/n$. Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1), $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$, and $a_n/h_n^{(p-q)/2} = O(1)$,*

$$\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n(\mathbf{V}) - L(\mathbf{V})| \to 0 \quad \textit{in probability as } n \to \infty \tag{20}$$

### 3.2 The Conditional Variance Estimator

Next we define the estimator we propose for $\text{span}\{\mathbf{B}\}$ in (1). Our main theoretical result follows in Theorem 4 which establishes the consistency of our estimator.

**Definition.** *The sample based* **Conditional Variance Estimator** $\widehat{B}_{p-q}$ *is any basis of* $\text{span}\{\widehat{\mathbf{V}}_q\}^\perp$ *where* $\widehat{\mathbf{V}}_q = \text{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L_n(\mathbf{V})$.

**Theorem 4.** *Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1),* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$*, and* $a_n/h_n^{(p-q)/2} = O(1)$*,* $\text{span}\{\widehat{\mathbf{B}}_k\}$ *is a consistent estimator for* $\text{span}\{\mathbf{B}\}$ *in model* (1)*; i.e.,*

$$\|\mathbf{P}_{\widehat{\mathbf{B}}_k} - \mathbf{P}_{\mathbf{B}}\| \to 0 \quad \textit{in probability as } n \to \infty.$$

### 3.3 Weighted estimation of $L(\mathbf{V})$

The set of points $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \le h_n\}$ represents a *slice* in the a subspace of $\mathbb{R}^p$ about $\mathbf{s}_0+\text{span}\{\mathbf{V}\}$. In the estimation of $L(\mathbf{V})$ two different weighting schemes are used:

(a) *Within a slice*. The weights are defined in (17) and are used to calculate (18).

(b) *Between slices*. Equal weights $1/n$ are used to calculate (19).

The choice of weights can be potentially influential. Especially the between weighting scheme can further be refined by assigning more weight to slices with more points. This can be realized by altering (19) to

$$L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^n \tilde{w}(\mathbf{V}, \mathbf{X}_i)\tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \quad \text{with} \tag{21}$$

$$\tilde{w}(\mathbf{V}, \mathbf{X}_i) = \frac{\sum_{j=1}^n K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n) - 1}{\sum_{l,u=1}^n K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n) - n} = \frac{\sum_{j=1, j\neq i}^n K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)}{\sum_{l,u=1, l\neq u}^n K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n)} \tag{22}$$

For example, if a rectangular kernel is used, $\sum_{j=1,j\neq i}^n K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$ is the number of $\mathbf{X}_j$ ($j \neq i$) points in the slice corresponding to $\tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$. Therefore this slice gets higher weight, if the number of $\mathbf{X}_j$ points in this slice is larger. That is, the more observations we use for estimating $L(\mathbf{V}, \mathbf{X}_i)$ the better its accuracy. The denominator in (22) guarantees the weights $\tilde{w}(\mathbf{V}, \mathbf{X}_i)$ sum up to one.

### 3.4 Bandwidth selection

The performance of conditional variance estimation depends crucially on the choice of the bandwidth sequence $h_n$ that controls the bias-variance trade-off if the mean squared error is used as measure for accuracy, in the sense that the smaller $h_n$ is, the lower the bias and the higher the variance and vice versa. Furthermore, the choice of $h_n$ depends on $p$, $q$, the sample size $n$, and the distribution of $\mathbf{X}$. We assume throughout the bandwidth satisfies assumptions (H.1) and (H.2). We will use Lemma 5 to derive a data-driven bandwidth we use in the computation of our estimator.

**Lemma 5.** *Let* $\mathbf{M}$ *be a* $p \times p$ *positive definite matrix. Then,*

$$\frac{tr(\mathbf{M})}{p} = \text{argmin}_{s>0} \|\mathbf{M} - s\mathbf{I}_p\| \tag{23}$$

*Proof.* Let $\mathbf{U}$ be the $p \times p$ matrix whose columns are the eigenvectors of $\mathbf{M}$ corresponding to its eigenvalues $\lambda_1 \ge \ldots \ge \lambda_p > 0$. Then, $\mathbf{M} = \mathbf{U}\text{diag}(\lambda_1, ..., \lambda_p)\mathbf{U}^T$, which implies $\|\mathbf{M} - s\mathbf{I}_p\|_2^2 = \|\text{diag}(\lambda_1, ..., \lambda_p) - s\mathbf{I}_p\|^2 = \sum_{l=1}^p (\lambda_l - s)^2$. Taking the derivative with respect to $s$, setting it to 0 and solving for $s$ obtains (23), since $\sum_{l=1}^p \lambda_l = \text{tr}(\mathbf{M})$. □

If the predictors are multivariate normal, their joint density is approximated by $N(\mu_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$ by Lemma 5, with $\sigma^2 = \text{tr}(\mathbf{\Sigma}_{\mathbf{x}})/p$. This results in no bandwidth dependence on $\mathbf{V}$ and leads to a rule for bandwidth selection, as follows.

Under $\mathbf{X} \sim N_p(\mu_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$, $\widetilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbf{X}_j \sim N_p(0, 2\sigma^2\mathbf{I}_p)$ for $i \neq j$, where we suppress the dependence on $j$ for notational convenience. Since all data are used as shifting points, $d_i(\mathbf{V}, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|^2 - (\mathbf{X}_i - \mathbf{X}_j)^T\mathbf{V}\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_j) = \|\widetilde{\mathbf{X}}_i\|^2 - \widetilde{\mathbf{X}}_i^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}}_i$. Let

$$\begin{aligned} \text{nObs} &= \mathbb{E}\left(\#\{i \in \{1, ..., n\} : \widetilde{\mathbf{X}}_i \in \text{span}_h\{\mathbf{V}\}\}\right) \\ &= 1 + (n-1)\mathbb{P}(d_1(\mathbf{V}, \mathbf{X}_2) \le h) = 1 + (n-1)\mathbb{P}(\|\widetilde{\mathbf{X}}\|^2 - \widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} \le h) \end{aligned} \tag{24}$$

where $\text{span}_h\{\mathbf{V}\} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h\}$ and $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{X}^*$, with $\mathbf{X}^*$ an independent copy of $\mathbf{X}$. nObs is the expected number of points in a slice. Given a user specified value for nObs, $h$ is the solution to (24).

Let $\mathbf{x} \in \mathbb{R}^p$. For any $\mathbf{V} \in \mathcal{S}(p,q)$ in (3), there exists an orthonormal basis $\mathbf{U} \in \mathbb{R}^{p\times(p-q)}$ of $\text{span}\{\mathbf{V}\}^\perp$ such that $\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$, by (5). Then, $\widetilde{\mathbf{X}} = \mathbf{V}\mathbf{R}_1 + \mathbf{U}\mathbf{R}_2$, with $\mathbf{R}_1 = \mathbf{V}^T\widetilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_q)$, $\mathbf{R}_2 = \mathbf{U}^T\widetilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_{p-q})$, and $\widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} = \|\mathbf{R}_1\|^2$ and $\|\widetilde{\mathbf{X}}\|^2 = \|\mathbf{R}_1\|^2 + \|\mathbf{R}_2\|^2$. Therefore,

$$\mathbb{P}\left(\|\widetilde{\mathbf{X}}\|^2 - \widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} \leq h\right) = \mathbb{P}(\|\mathbf{R}_2\|^2 \leq h) = \chi_{p-q}\left(\frac{h}{2\sigma^2}\right), \tag{25}$$

where $\chi_{p-q}$ is the cumulative distribution function of a chi-squared random variable with $p-q$ degrees of freedom. Plugging (25) in (24) obtains

$$\text{nObs} = 1 + (n-1)\chi_{p-q}\left(\frac{h}{2\sigma^2}\right). \tag{26}$$

Solving (26) for $h$ and Lemma 5 yield

$$h_n(\text{nObs}) = \chi^{-1}_{p-q}\left(\frac{\text{nObs}-1}{n-1}\right)\frac{2\text{tr}(\widehat{\Sigma}_\mathbf{x})}{p}, \tag{27}$$

where $\widehat{\Sigma}_\mathbf{x} = \sum_i(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/n$ and $\bar{\mathbf{X}} = \sum_i \mathbf{X}_i/n$.

In order to ascertain $h_n$ satisfies (H.1) and (H.2), a reasonable choice is to set nObs $= \gamma(n)$ for a function $\gamma(\cdot)$ with $\gamma(n) \to \infty$, $\gamma(n)/n \leq 1$ and $\gamma(n)/n \to 0$. For example, nObs $= \gamma(n) = n^\beta$ with $\beta \in (0,1)$ can be used.

Alternatively, a plug-in bandwidth based on rule-of-thumb rules of the form $csn^{-1/(4+k)}$, where $s$ is an estimate of scale and $c$ a number close to 1, such as Silverman's ($c = 1.06$, $s$ =standard deviation) or Scott's ($c = 1$, $s$ =standard deviation), used in nonparametric density estimation [see [29]], is

$$h_n = 1.2^2\frac{2\text{tr}(\widehat{\Sigma}_\mathbf{x})}{p}\left(n^{-1/(4+p-q)}\right)^2. \tag{28}$$

The term $2\text{tr}(\widehat{\Sigma}_\mathbf{X})/p$ can be interpreted as the variance of $\mathbf{X}_i - \mathbf{X}_j$ and $p-q$ is the true dimension $k$. We use 1.2 as $c$ based on empirical evidence from simulations. Since both (27) and (28) yield satisfactory results, we opted against cross validation for bandwidth selection because of the computational burden involved, and used the bandwidth in (28) in simulations and data analyses.

## 4 Optimization Algorithm

A Stiefel manifold optimization algorithm is used to obtain the solution of the sample version of the optimization problem (13). To calculate $\widehat{\mathbf{V}}_q$ in (3.2), a curvilinear search is carried out [33, 31], which is similar to gradient descent. First an arbitrary starting value $\mathbf{V}^{(0)}$ is selected by drawing a $p \times q$ matrix from the invariant measure; i.e., the distribution that corresponds to the uniform, on $\mathcal{S}(p,q)$, see [8]. The $Q$-component of the QR decomposition of a $p \times q$ matrix with independent standard normal entries follows the invariant measure [7]. The step-size $\tau > 0$, the step size reduction factor $\gamma \in (0,1)$, and tolerance tol $> 0$ are fixed at the outset.

**Result:** $\mathbf{V}^{(\text{end})}$
Initialize: $\mathbf{V}^{(0)}$, $\tau = 1$, tol $= 10^{-3}$, $\gamma = 0.5$ error $=$ tol $+ 1$, maxit $= 50$, count $= 0$;
**while** *error* $>$ *tol and count* $\leq$ *maxit* **do**

- $\mathbf{G} = \nabla_\mathbf{V} L_n(\mathbf{V}^{(j)}) \in \mathbb{R}^{p\times q}$, $\mathbf{W} = \mathbf{G}\mathbf{V}^T - \mathbf{V}\mathbf{G}^T$
- $\mathbf{V}^{(j+1)} = (\mathbf{I}_p + \tau\mathbf{W})^{-1}(\mathbf{I}_p - \tau\mathbf{W})\mathbf{V}^{(j)}$
- error $= \|\mathbf{V}^{(j)}\mathbf{V}^{(j)\mathbf{T}} - \mathbf{V}^{(j+1)}\mathbf{V}^{(j+1)^T}\|/\sqrt{2q}$

  **if** $L_n(\mathbf{V}^{(j+1)}) > L_n(\mathbf{V}^{(j)})$ **then**
    $\mathbf{V}^{(j+1)} \leftarrow \mathbf{V}^{(j)}$; $\tau \leftarrow \tau\gamma$; error $\leftarrow$ tol $+ 1$
  **else**
    count $\leftarrow$ count $+ 1$
    $\tau \leftarrow \frac{\tau}{\gamma}$
  **end**

**end**

**Algorithm 1:** Curvilinear search

Under mild regularity conditions on the objective function, [33] showed that the sequence generated by the algorithm converges to a stationary point if the Armijo-Wolfe conditions [27] are used for determining the stepsize $\tau$.

The Armijo-Wolfe conditions require the evaluation of the gradient for each potential step size until one is found that fulfills the conditions and the step is accepted, i.e. for the determination of one step size the gradient has to be evaluated multiple times. Since for the conditional variance estimator, the gradient computation incurs the highest computational cost, we use simpler conditions to determine the step size. Specifically, we simply require the step decrease the objective function, otherwise the step size $\tau$ is decreased by the factor $\gamma \in (0, 1)$). These simplified conditions are computationally less expensive and exhibit same behavior as the Armijo-Wolfe conditions in the simulations. Further we capped the maximum number of steps at maxit $= 50$ steps, since the algorithm converged in about 10 iterations in all our simulations.

The algorithm is repeated for $m$ arbitrary $\mathbf{V}^{(0)}$ starting values drawn from the invariant measure on $\mathcal{S}(p, q)$. Among those, the value at which $L_n$ in (19) is minimal is selected as $\widehat{\mathbf{V}}_q$.

The algorithm requires the computation of the gradient of $L_n(\mathbf{V})$ in (19) or (21). We compute the gradient of the objective function for the Gaussian kernel in Theorems 6 and 7. The Gaussian kernel is the default kernel we use in the implementation of the estimation algorithm in the `R` code that accompanies this manuscript.

**Theorem 6.** *Let $K(z) = \exp(-z^2/2)$ be the Gaussian kernel. Then, the gradient of $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ in* (18) *is given by*

$$\nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{s}_0) = \frac{1}{h_n^2} \sum_{i=1}^{n} (\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) - (Y_i - \bar{y}_1(\mathbf{V}, \mathbf{s}_0))^2) w_i d_i \nabla_{\mathbf{V}} d_i(\mathbf{V}, \mathbf{s}_0) \in \mathbb{R}^{p \times q},$$

*and the gradient of $L_n(\mathbf{V})$ in* (19) *is*

$$\nabla_{\mathbf{V}} L_n(\mathbf{V}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i).$$

*with $w_i = w(\mathbf{V}, \mathbf{X}_i)$ in* (17)*.*

The weighted version of conditional variance estimation in Section 3.3 is expected to increase the accuracy of the estimator for unevenly spaced data. When (21) and the gradient in (29) are used in the optimisation algorithm, we refer to the estimator as *weighted conditional variance estimation*. If (21) and the gradient $\sum_{i=1}^{n} \tilde{w}(\mathbf{V}, \mathbf{X}_i) \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$ is used; i.e., the first summand in (29) is dropped, we refer to it as *partially weighted conditional variance estimation*. For both, we replace $G$ in algorithm 1 with the corresponding gradient derived in Theorem 7.

**Theorem 7.** *Let $K(z) = \exp(-z^2/2)$ be the Gaussian kernel. Then, the gradient of $L_n^{(w)}(\mathbf{V})$ in* (21) *is given by*

$$\nabla_{\mathbf{V}} L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^{n} \left( \nabla_{\mathbf{V}} \tilde{w}(\mathbf{V}, \mathbf{X}_i) \tilde{L}_n(\mathbf{V}, \mathbf{X}_i) + \tilde{w}(\mathbf{V}, \mathbf{X}_i) \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i) \right), \tag{29}$$

*where $\nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$ is given in Theorem 6. Furthermore,*

$$\nabla_{\mathbf{V}} \tilde{w}(\mathbf{V}, \mathbf{X}_i) = -\frac{1}{h_n^2} \sum_{j} \left( \frac{K_{j,i}}{\sum_{l,u=1}^{n} K_{l,u}} d_{j,i} \nabla_{\mathbf{V}} d_{j,i} - \tilde{w}_i \sum_{l,u=1}^{n} \frac{K_{l,u}}{\sum_{o,s=1}^{n} K_{o,s}} d_{l,u} \nabla_{\mathbf{V}} d_{l,u} \right)$$

*with $\tilde{w}_i = \tilde{w}(\mathbf{V}, \mathbf{X}_i)$ in* (22)*, $K_{j,i} = K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$, and $d_{j,i} = d_j(\mathbf{V}, \mathbf{X}_i)$ given in* (16)*.*

### 4.1 A study of the behaviour of $L_n(\mathbf{V})$

We explore how accurately the sample version (19) of the objective function estimates the target subspace in an example. We consider a bivariate normal predictor vector, $\mathbf{X} = (X_1, X_2)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}})$. We generate the response from $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon = X_1 + \epsilon$, with $\epsilon \sim N(0, \eta^2)$ independent of $\mathbf{X}$. In this setting, $k = 1, \mathbf{B} = (1, 0)^T$, $g(z) = z \in \mathbb{R}$ in model (1). With these specifications, (10) becomes

$$\mu_l(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}} (\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}r)^l f_{\mathbf{X}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}}(r) dr \tag{30}$$

Dropping the terms that do not contain $\mathbf{r}$ in (8) yields

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(r) \propto f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}r) \propto \exp\left(-\frac{1}{2}(\mathbf{s}_0+r\mathbf{V})^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{s}_0+r\mathbf{V})\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(2r\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0+r^2\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}\right)\right) = \exp\left(-\frac{1}{2\sigma^2}\left(2r\sigma^2\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0+r^2\right)\right)$$
$$\propto \exp\left(-\frac{1}{2\sigma^2}(r-\alpha)^2\right), \tag{31}$$

where $\sigma^2 = 1/(\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V})$, $\alpha = -\sigma^2\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0$ and the symbol $\propto$ stands for proportional to. Letting $\psi(z)$ denote the density of a standard normal variable, (31) obtains

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(r) = \frac{1}{\sigma}\psi\left(\frac{r-\alpha}{\sigma}\right) \tag{32}$$

for $\mathbf{V}, \mathbf{s}_0 \in \mathbb{R}^{2\times 1}$. Inserting (32) in (30) yields

$$\int_{\mathbb{R}}(\mathbf{B}^T\mathbf{s}_0+\mathbf{B}^T\mathbf{V}r)^l\frac{1}{\sigma}\psi\left(\frac{r-\alpha}{\sigma}\right)dr = \begin{cases}\mathbf{B}^T\mathbf{s}_0+\mathbf{B}^T\mathbf{V}\alpha & l=1\\ (\mathbf{B}^T\mathbf{s}_0)^2+2(\mathbf{B}^T\mathbf{s}_0)(\mathbf{B}^T\mathbf{V})\alpha+(\mathbf{B}^T\mathbf{V})^2(\sigma^2+\alpha^2) & l=2\end{cases}$$

Using (9), (6) and (7), yields $\tilde{L}(\mathbf{V},\mathbf{s}_0) = \mu_2(\mathbf{V},\mathbf{s}_0) - \mu_1(\mathbf{V},\mathbf{s}_0)^2 + \eta^2 = (\mathbf{B}^T\mathbf{V})^2\sigma^2 + \eta^2$, so that

$$L(\mathbf{V}) = \mathbb{E}\left(\tilde{L}(\mathbf{V},\mathbf{X})\right) = (\mathbf{B}^T\mathbf{V})^2\sigma^2+\eta^2 = \frac{(\mathbf{B}^T\mathbf{V})^2}{\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}}+\eta^2 \tag{33}$$

From (33) we can easily see that $L(\mathbf{V})$ attains its minimum at $\mathbf{V}\perp\mathbf{B}$. Also, if $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{I}_2$, the maximum of $L(\mathbf{V})$ is attained at $\mathbf{V} = \mathbf{B}$. To visualize the behavior of $\tilde{L}_n(\mathbf{V})$ as the sample size increases, we parametrize $\mathbf{V}$ by $\mathbf{V}(\theta) = (\cos(\theta),\sin(\theta))^T$, $\theta\in[0,\pi]$. Since $\mathbf{B} = (1,0)^T$, the minimum of $\tilde{L}(\mathbf{V})$ is at $\mathbf{V}(\pi/2) = (0,1)^T$ , which is orthogonal to $\mathbf{B}$.

The true $L(\mathbf{V}(\theta))$ and its estimates $L_n(\mathbf{V}(\theta))$ are plotted for samples of different sizes $n$ in Figure 1. $L_n(\mathbf{V}(\theta))$ approximates $L(\mathbf{V})$ fast and attains its minimum at the same value as $L(\mathbf{V})$ even for $n = 10$.

As an aside, we note that assumption (A.4) is violated in this example, which suggests that the proposed estimator of conditional variance estimation may apply under weaker assumptions.

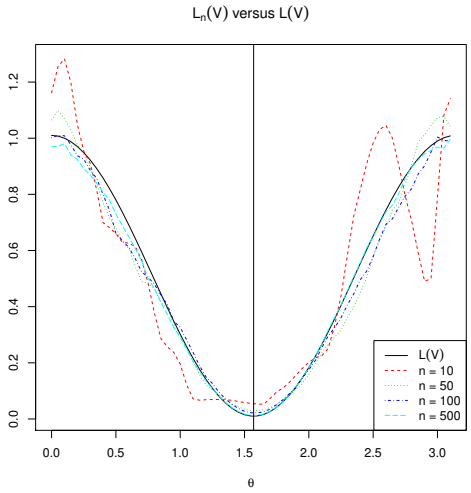


Figure 1: Solid black line is $L(\mathbf{V}(\theta)) = \cos(\theta)^2 + 0.1^2$, colored is $L_n(\mathbf{V}(\theta))$, $\theta\in[0,\pi]$, $n = 10, 50, 100, 500$. The vertical black line is at $\theta = \pi/2$

# 5 Simulation studies

We compare the estimation accuracy of conditional variance estimation with the forward model based sufficient dimension reduction methods, mean outer product gradient estimation (`meanOPG`), mean minimum average variance estimation (`meanMAVE`) [32], refined outer product gradient (`rOPG`), refined minimum average variance estimation (`rmave`) [35, 22], and principal Hessian directions (`pHd`) [24, 15], and the inverse regression based methods, sliced inverse regression (`SIR`) [23] and sliced average variance estimation (`SAVE`) [14]. The dimension $k$ is assumed to be known throughout.

We report results for conditional variance estimation using the "plug-in" bandwidth in (28) and three different conditional variance estimation versions, `CVE`, `wCVE`, and `rCVE`. `CVE` is obtained by using $m = 10$ arbitrary starting values in the optimization algorithm and optimizing (19) as described in Section 4. `rCVE`, or *refined weighted CVE*, is obtained by setting the starting value $\mathbf{V}^{(0)}$ at the optimizer of `CVE`, and using (21) in the optimization algorithm in Section 4 with the partially weighted gradient as described in Section 3.3. `wCVE`, or *weighted CVE*, is obtained by optimizing (21) with partially weighted gradient as described in Sections 3.3 and 4. Methods `rOPG` and `rmave` refer to the original refined outer product gradient and refined minimum average variance estimation algorithms published in [35]. They are implemented using the `R` code in [22] with number of iterations nit $= 25$, since the algorithm is seen to converge by 25. The `dr` package is used for the `SIR`, `SAVE` and `pHd` calculations, and the `MAVE` package for mean outer product gradient estimation (`meanOPG`) and mean minimum average variance estimation (`meanMAVE`). The source code for conditional variance estimation can be downloaded from `https://git.art-ist.cc/daniel/CVE`.

Table 1 lists the seven models (M1-M7) we consider. Throughout, we set $p = 20$, $\mathbf{b}_1 = (1, 1, 1, 1, 1, 1, 0, ..., 0)^T/\sqrt{6}$, $\mathbf{b}_2 = (1, -1, 1, -1, 1, -1, 0, ..., 0)^T/\sqrt{6} \in \mathbb{R}^p$ for M1-M5. For M6, $\mathbf{b}_1 = \mathbf{e}_1, \mathbf{b}_2 = \mathbf{e}_2$ and $\mathbf{b}_3 = \mathbf{e}_p$, and for M7 $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are the same as in M6 and $\mathbf{b}_4 = \mathbf{e}_3$, where $\mathbf{e}_j$ denotes the $p$-vector with $j$th element equal to 1 and all others are 0. The error term $\epsilon$ is independent of $\mathbf{X}$ for all models. In M2, M3, M4, M5 and M6, $\epsilon \sim N(0, 1)$. For M1 and M7, $\epsilon$ has a generalized normal distribution $GN(a, b, c)$ with densitiy $f_\epsilon(z) = c/(2b\Gamma(1/c)) \exp((|z - a|/b)^c)$, see [26] with location 0 and shape-parameter 0.5 for M1, and shape-parameter 1 for M7 (Laplace distribution). For both the scale-parameter is chosen such that $\mathbb{V}\mathrm{ar}(\epsilon) = 0.25$.

Table 1: Models

| Name | Model | $\mathbf{X}$ distribution | $\epsilon$ distribution | $k$ | $n$ |
|---|---|---|---|---|---|
| M1 | $Y = \cos(\mathbf{b}_1^T\mathbf{X}) + \epsilon$ | $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$ | $GN(0, \sqrt{1/2}, 0.5)$ | 1 | 100 |
| M2 | $Y = \cos(\mathbf{b}_1^T\mathbf{X}) + 0.5\epsilon$ | $\mathbf{X} \sim \lambda Z\mathbf{1}_p + N_p(\mathbf{0}, \mathbf{I}_p)$ | $N(0, 1)$ | 1 | 100 |
| M3 | $Y = 2\log(\lvert\mathbf{b}_1^T\mathbf{X}\rvert + 2) + 0.5\epsilon$ | $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ | $N(0, 1)$ | 1 | 100 |
| M4 | $Y = (\mathbf{b}_1^T\mathbf{X})/(0.5 + (1.5 + \mathbf{b}_2^T\mathbf{X})^2) + 0.5\epsilon$ | $\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$ | $N(0, 1)$ | 2 | 200 |
| M5 | $Y = \cos(\pi\mathbf{b}_1^T\mathbf{X})(\mathbf{b}_2^T\mathbf{X} + 1)^2 + 0.5\epsilon$ | $\mathbf{X} \sim U([0, 1]^p)$ | $N(0, 1)$ | 2 | 200 |
| M6 | $Y = (\mathbf{b}_1^T\mathbf{X})^2 + (\mathbf{b}_2^T\mathbf{X})^2 + (\mathbf{b}_3^T\mathbf{X})^2 + 0.5\epsilon$ | $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$ | $N(0, 1)$ | 3 | 200 |
| M7 | $Y = (\mathbf{b}_1^T\mathbf{X})(\mathbf{b}_2^T\mathbf{X})^2 + (\mathbf{b}_3^T\mathbf{X})(\mathbf{b}_4^T\mathbf{X}) + \epsilon$ | $\mathbf{X} \sim t_3(\mathbf{I}_p)$ | $GN(0, \sqrt{1/\Gamma(6)}, 1)$ | 4 | 400 |

The variance-covariance structure of $\mathbf{X}$ in models M1 and M4 satisfies $\mathbf{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$. In M5, $\mathbf{X}$ is uniform with independent entries on the $p$-dimensional hyper-cube. In M7, $\mathbf{X}$ is multivariate $t$-distributed with 3 degrees of freedom. The link functions of M4 and M7 are studied in [35], but we use $p = 20$ instead of 10 and a non identity covariance structure for M4 and the $t$-distribution instead of normal for M7. In M2, $Z \sim 2\text{Bernoulli}(p_{\text{mix}}) - 1 \in \{-1, 1\}$, where $\mathbf{1}_q = (1, 1, ..., 1)^T \in \mathbb{R}^q$, mixing probability $p_{\text{mix}} \in [0, 1]$ and dispersion parameter $\lambda > 0$. For $0 < p_{\text{mix}} < 1$, $\mathbf{X}$ has a mixture normal distribution, where $p_{\text{mix}}$ is the relative mode height and $\lambda$ is a measure of mode distance.

We set $q = p - k$ and generate $r = 100$ replications of models M1 - M7. We estimate $\mathbf{B}$ using the ten sufficient dimension reduction methods. The accuracy of the estimates is assessed using $err = \|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|/\sqrt{2k}$, which lies in the interval $[0, 1]$. The factor $\sqrt{2k}$ normalizes the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement , specifically, $\|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|^2 \leq 2k$.

In Table 2 the mean and standard deviation of $err$ for M1 - M7 are reported. In particular, for M2, $p_{mix} = 0.3$ and $\lambda = 1$. The smallest error values are boldfaced. In models M1, M2 and M3, the conditional variance estimator is the best performer, with its refined version as close second. In M4, M5 and M6, any of the four versions of MAVE performs better than the CVE. For model M7 the results of `rOPG` and `rmave` are not reported because the code frequently produces an error message that a matrix is not invertible. Among the rest, the weighted version of CVE, wCVE, attains the minimum error.

Sliced inverse regression (SIR) and sliced average variance estimation (SAVE) are not competitive throughout our experiments. Sliced inverse regression (SIR), in particular, is expected to fail in models M1-M3, and M6 since $\mathbb{E}(Y \mid \mathbf{X})$ is even.

In Figure 2, box-plots for all combinations of $p_{\text{mix}} \in \{0.3, 0.4, 0.5\}$ and $\lambda \in \{0, 0.5, 1, 1.5\}$ are presented. The reference methods are restricted to meanOPG and meanMAVE, since the others are not competitive. Conditional variance estimation performs better than all competing methods and is the only method with consistently smaller errors when the two modes are further apart ($\lambda \geq 1$) regardless of the mixing probability $p_{\text{mix}}$. The performance of both meanOPG and meanMAVE worsens as one moves from left to right row-wise. The mixing probability, $p_{\text{mix}}$, has no noticeable effect on the performance of any method; i.e., the plots are very similar column-wise. In sum, meanMAVE's performance deteriorates as the bimodality of the predictor distribution becomes more distinct. In contrast, conditional variance estimation is unaffected. and appears to have an advantage over meanMAVE when the predictors have mixture distributions, the link function is even about the midpoint of the two modes, and $\mathbf{B}$ is not orthogonal to the line connecting the two modes. Conditional variance estimation is the only method that estimates the mean subspace reliably in model M2 ($err \approx 0.4$ to $0.5$), whereas meanMAVE misses it completely ($err \approx 1$). These results indicate that conditional variance estimation is often approximately on par, and can perform much better than meanMAVE depending on the predictor distribution and the link function.

Table 2: Mean and standard deviation of estimation errors

| Model | | CVE | wCVE | rCVE | meanOPG | rOPG | meanMAVE | rmave | pHd | sir | save |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $M1$ | | | | | | | | | | | |
| | mean | **0.3827** | 0.4414 | 0.4051 | 0.6220 | 0.9876 | 0.5099 | 0.9840 | 0.8278 | 0.9875 | 0.9788 |
| | sd | 0.1269 | 0.1595 | 0.1329 | 0.1879 | 0.0223 | 0.1800 | 0.0295 | 0.1206 | 0.0243 | 0.0334 |
| $M2$ | | | | | | | | | | | |
| | mean | **0.4572** | 0.4992 | 0.4658 | 0.8987 | 0.9332 | 0.8905 | 0.9242 | 0.9000 | 0.9783 | 0.9781 |
| | sd | 0.1038 | 0.1524 | 0.0989 | 0.0908 | 0.0683 | 0.0983 | 0.0897 | 0.0735 | 0.0278 | 0.0318 |
| $M3$ | | | | | | | | | | | |
| | mean | **0.6282** | 0.7509 | 0.6371 | 0.7847 | 0.9644 | 0.7576 | 0.9674 | 0.6964 | 0.9647 | 0.9519 |
| | sd | 0.2354 | 0.2262 | 0.2181 | 0.2201 | 0.0667 | 0.2435 | 0.0609 | 0.1626 | 0.0587 | 0.0650 |
| $M4$ | | | | | | | | | | | |
| | mean | 0.5663 | 0.5897 | 0.5554 | 0.4071 | 0.4026 | 0.4361 | **0.3905** | 0.7772 | 0.5824 | 0.9727 |
| | sd | 0.1239 | 0.1246 | 0.1298 | 0.0814 | 0.0609 | 0.0997 | 0.0584 | 0.0662 | 0.0951 | 0.0202 |
| $M5$ | | | | | | | | | | | |
| | mean | 0.4429 | 0.5604 | 0.4779 | 0.4058 | **0.3737** | 0.3929 | 0.3750 | 0.7329 | 0.6374 | 0.9730 |
| | sd | 0.0891 | 0.1233 | 0.0976 | 0.1022 | 0.0680 | 0.0894 | 0.0871 | 0.0832 | 0.0968 | 0.0186 |
| $M6$ | | | | | | | | | | | |
| | mean | 0.3828 | 0.3027 | 0.3230 | 0.1827 | 0.4632 | **0.1656** | 0.4863 | 0.4978 | 0.9129 | 0.8236 |
| | sd | 0.1006 | 0.0748 | 0.1098 | 0.0289 | 0.1717 | 0.0252 | 0.1676 | 0.0601 | 0.0420 | 0.0518 |
| $M7$ | | | | | | | | | | | |
| | mean | 0.6856 | **0.5050** | 0.5651 | 0.5694 | NA | 0.5482 | NA | 0.8536 | 0.8133 | 0.8699 |
| | sd | 0.0588 | 0.0862 | 0.0879 | 0.1122 | NA | 0.1271 | NA | 0.0354 | 0.0341 | 0.0342 |

Furthermore we estimate the dimension $k$ via cross-validation, following the approach in [35] , with

$$\hat{k} = \text{argmin}_{l=1,\ldots,p}\, CV(l) = \text{argmin}_{l=1,\ldots,p}\, \frac{\sum_i (Y_i - \hat{g}^{-i}(\widehat{\mathbf{B}}_l^T \mathbf{X}_i))^2}{n}, \tag{34}$$

where $\hat{g}^{-i}(\cdot)$ is computed from the data $(Y_j, \widehat{\mathbf{B}}_l^T \mathbf{X}_j)_{j=1,\ldots,n; j\neq i}$ using multivariate adaptive regression splines [17] in the R-package mda, and $\widehat{\mathbf{B}}_l = \widehat{\mathbf{V}}_{p-l}^{\perp}$ is any basis of the orthogonal complement of $\widehat{\mathbf{V}}_{p-l} = \text{argmin}_{\mathbf{V}\in\mathcal{S}(p,p-l)}\, L_n(\mathbf{V})$. For a given $l$, we calculate $\widehat{\mathbf{B}}_l$ from the whole data set and predict $Y_i$ by $\hat{Y}_{i,l} = \hat{g}^{-i}(\widehat{\mathbf{B}}_l^T \mathbf{X}_i)$. For $l = p$, $\widehat{\mathbf{B}}_p = \mathbf{I}_p$. The results for the seven models are reported in Table 3. The CVE based dimension estimation is the most accurate in models M1, M2, M3, and M6 and differs slightly from that of MAVE in M7. MAVE performs better in M4 and M5, completely misses the true dimension in M2 and misses it most of the time in M3. Thus, the dimension estimation performance of CVE and MAVE agrees with the estimation accuracy of the true subspace in Table 2, CVE estimates the dimension more accurately even in model M6, where it exhibits worse subspace estimation performance, and overall appears to be more accurate.

We carried out many simulation experiments for an array of combinations of link functions, sufficient reduction matrices $\mathbf{B}$ and their ranks, as well as predictor and error distributions. All reported and unreported results indicate that the

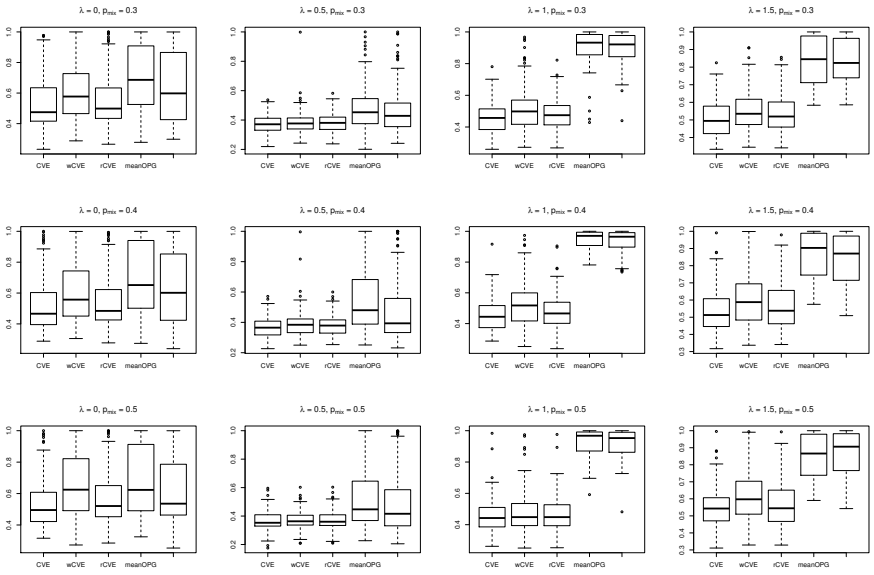


Figure 2: M2, $p = 20, n = 100$

Table 3: Number of times dimension $k$ is correctly estimated in 100 replications

| | M1 | M2 | M3 | M4 | M5 | M6 | M7 |
|---|---|---|---|---|---|---|---|
| CVE | 83 | 41 | 88 | 62 | 46 | 74 | 19 |
| MAVE | 67 | 0 | 14 | 76 | 60 | 57 | 21 |

difference in performance of the two methods, CVE and mean MAVE, can be attributed to both the form of the link function and the marginal predictor distribution. We observed that when the link function had a bounded first order derivative, CVE often outperformed mean MAVE across predictor distributions. In the opposite case, MAVE performed mostly better. Also, when the predictors have a bimodal distribution with well separated modes and the link function is even, regardless of whether its derivative is bounded, CVE outperforms mean MAVE. In the other settings for the generated data, both methods were roughly on par.

## 6 Real Data Analyses

Three data sets are analyzed: the *Hitters* data in the `R` package `ISLR`, which was also analyzed by [35], the *Boston Housing* data in the `R` package `mlbench`, and the *Concrete* data from the `MAVE` package. The reference method is `meanMAVE` from the `MAVE` package in `R` and the `CVE` is calculated using $m = 50$ and maxit $= 10$ in the optimization algorithm 1 in Section 4. The estimation of the dimension is based on (34) in Section 5.

Following [35], we remove 7 outliers from the *Hitters* data set leading to a sample size of 256. The response is $Y = \log(\text{salary})$ and the 16 continuous predictors are the game statistics of players in the Major League Baseball league in the seasons 1986 and 1987. Further information can be found in `https://www.rdocumentation.org/packages/ISLR/versions/1.2/topics/Hitters`.

The *Boston Housing* data set contains 506 census tracts on 14 variables from the 1970 census. The response is `medv`, the median value of owner-occupied homes in USD 1000's. The factor variable `chas` is removed from the data set for the analysis so that the response is modeled by the remaining 12 continous predictors. The description of the variables can be found in `https://www.rdocumentation.org/packages/mlbench/versions/2.1-1/topics/BostonHousing`.

The *Concrete* data set contains 1030 instances on 9 continuous variables The response is concrete compressive strength. Concrete strength is very important in civil engineering and is a highly nonlinear function of age and ingredients. The

description of the variables can be found in `https://www.rdocumentation.org/packages/MAVE/versions/1.3.10/topics/Concrete`.

For all three data sets we standardize both the predictors and the response by subtracting the mean and rescaling column-wise so that each variable has unit variance. The data sets are analyzed using 10 fold cross-validation to calculate an unbiased estimate of the prediction error [30] for our method , `CVE`, and its main competitor `meanMAVE` using the `MAVE` package. The dimension for each method is estimated with (34) on the trainings set and we then fit a forward regression model on the training set replacing the original with the reduced predictors using multivariate adaptive regression splines [17] using the `R` package `mda` and calculate the prediction error on the test set for both methods. The dimension estimates of `CVE` and `MAVE` mostly disagree.

The mean and standard deviation of the 10-fold cross-validation prediction errors are reported in Table 4. Since the response is standardized, the values in Table 4 are bounded between 0 and 1, with smaller values indicating better predictive performance. `CVE` performs slightly worse than mean `MAVE` in the *Hitters* data set, slightly better in the *Boston Housing* and better in the *Concrete* data set analysis.

Table 4: Mean and standard deviation (in parenthesis) of standardized out of sample prediction errors for the three data sets

| Method | Hitters | Housing | Concrete |
|---|---|---|---|
| CVE | 0.216 | 0.260 | 0.361 |
| | (0.101) | (0.331) | (0.206) |
| MAVE | 0.203 | 0.299 | 0.417 |
| | (0.083) | (0.382) | (0.348) |

### 6.1 Hitters Data Analysis as in [35]

Additionally, we reconstruct the analysis of the *Hitters* data in [35], which does not account for the out-of-sample prediction error as in Section 6 but uses the whole sample for estimation of $\mathbf{B}$ and its rank. Only the dimension $k$ is estimated with leave-one-out cross validation.

Table 5 reports the average cross validation mean squared error $CV(k)$ in (34) using the whole data set over $k = 1, \ldots, 5$. Both conditional variance estimation and mean minimum average variance estimation estimate the dimension to be 2.

Table 5: Mean cross-validation error

| $k$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CVE | 0.308 | 0.218 | 0.275 | 0.327 | 0.371 |
| MAVE | 0.370 | 0.277 | 0.339 | 0.413 | 0.440 |

We plot the response against the estimated directions in Figure 3. Both exhibit the same pattern: the response appears to be linear in one direction and quadratic in the second. The difference is that the linear pattern is clearer in the second CVE direction and the quadratic pattern exhibits increasing variance in the first MAVE direction.

Based on the scatterplots in Figure 3, we fit the same models for both. For conditional variance estimation, the fitted regression is

$$\hat{Y} = 0.39578 + 0.33724(\widehat{\mathbf{b}}_1^T\mathbf{X}) - 0.08066(\widehat{\mathbf{b}}_1^T\mathbf{X})^2 + 0.29126(\widehat{\mathbf{b}}_2^T\mathbf{X}) \tag{35}$$

with $R^2 = 0.7975$, and for minimum average variance estimation

$$\hat{Y} = 0.39051 + 1.32529(\widehat{\mathbf{b}}_1^T\mathbf{X}) - 0.55328(\widehat{\mathbf{b}}_1^T\mathbf{X})^2 + 0.49546(\widehat{\mathbf{b}}_2^T\mathbf{X}) \tag{36}$$

with $R^2 = 0.7859$. Both models (35) and (36) have about the same fit as measured by $R^2$. The in sample performance of the two methods is practically the same for the `Hitters` data.

## 7 Discussion

In this paper the novel conditional variance estimator (CVE) for the mean subspace is introduced. We present its geometrical and theoretical foundation, show its consistency and propose an estimation algorithm with assured

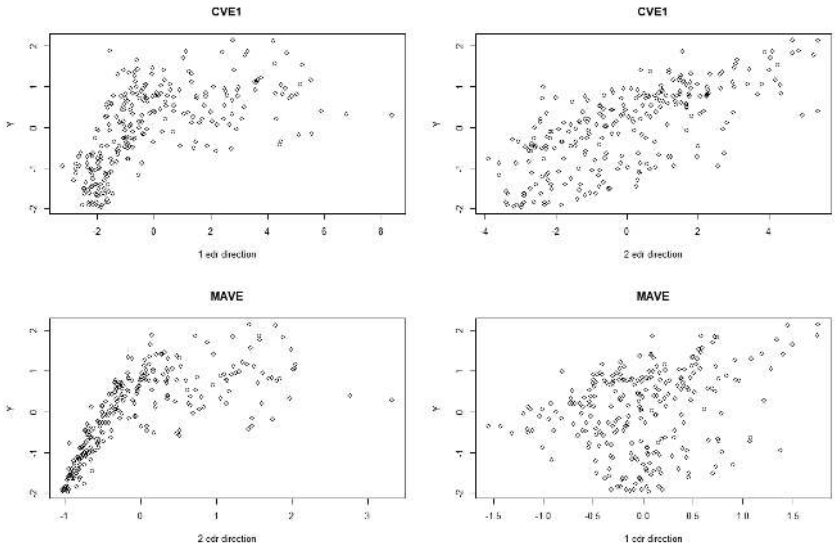


Figure 3: $Y$ against $\widehat{\mathbf{b}}_1^T\mathbf{X}$ and $\widehat{\mathbf{b}}_2^T\mathbf{X}$

convergence. CVE requires the forward model (1), $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$, holds and weak assumptions on the response and the covariates.

Minimum average variance estimation (MAVE) [35] is the only other sufficient dimension reduction method based on the forward model (1). It estimates the sufficient dimension reduction targeting both the reduction and the link function $g$ in (1). CVE targets only the reduction and does not require estimation of the link function, which may explain why it has an advantage over MAVE in some regression settings. For example, CVE exhibits similar performance across different link functions (cos, exp, etc) for fixed $\lambda$, whereas the performance of MAVE is very uneven for model M2 in Section 5. CVE is more accurate than MAVE when the link function is even and the predictor distribution is bimodal throughout our simulation studies. Moreover, CVE does not require the inversion of the predictor covariance matrix and can be applied to regressions with $p \approx n$ or $p > n$.

The theoretical challenge in deriving the statistical properties of conditional variance estimation arises from the novelty of its definition that involves random non i.i.d. weights that depend on the parameter to be estimated.

## References

[1] Kofi P. Adragni and R. Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 11 2009.

[2] Takeshi Amemiya. *Advanced Econometrics*. Harvard university press, 1985.

[3] S. N. Bernstein. *Theory of Probability*. Moscow, 1927.

[4] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2002.

[5] Efstathia Bura, Sabrina Duarte, and Liliana Forzani. Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, 111(515):1313–1329, 2016.

[6] Efstathia Bura and Liliana Forzani. Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, 110(509):420–434, 2015.

[7] Yasuko Chikuse. *Invariant measures on Stiefel manifolds with applications to multivariate analysis*, volume Volume 24 of *Lecture Notes–Monograph Series*, pages 177–193. Institute of Mathematical Statistics, Hayward, CA, 1994.

[8] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer-Verlag New York, New York, 2003.

[9] Dennis R. Cook. *Regression Graphics: Ideas for studying regressions through graphics*. Wiley, New York, 1998.

[10] R. D. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501, 2008.

[11] R. Dennis Cook. Fisher lecture: Dimension reduction in regression. *Statist. Sci.*, 22(1):1–26, 02 2007.

[12] R. Dennis Cook and Liliana Forzani. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 3 2009.

[13] R. Dennis Cook and Bing Li. Determining the dimension of iterative hessian transformation. *Ann. Statist.*, 32(6):2501–2531, 12 2004.

[14] R. Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[15] R.Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2):455–474, 04 2002.

[16] Arnold M. Faden. The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

[17] Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[18] Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748, 2008.

[19] H. Heuser. *Analysis 2, 9 Auflage*. Teubner, 1995.

[20] Alan F. Karr. *Probability*. Springer Texts in Statistics. Springer-Verlag New York, 1993.

[21] D. Leao Jr., M. Fragoso, and P. Ruffino. Regular conditional probability, disintegration of probability and radon spaces. *Proyecciones (Antofagasta)*, 23:15 – 29, 05 2004.

[22] Bing Li. *Sufficient dimension reduction: methods and applications with R*. CRC Press, Taylor & Francis Group, 2018.

[23] K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[24] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[25] Yanyuan Ma and Liping Zhu. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 4 2013.

[26] Saralees Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005.

[27] J. Nocedal and S. Wright. *Line Search Methods*, pages 30–65. Springer New York, New York, NY, 2006.

[28] E Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1961.

[29] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[30] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

[31] Hemant D. Tagare. Notes on optimization on stiefel manifolds, January 2011.

[32] Hang Weiqiang and Xia Yingcun. *MAVE: Methods for Dimension Reduction*, 2019. R package version 1.3.10.

[33] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142:397–434, 2013.

[34] Yingcun Xia. A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103(484):1631–1640, 2008.

[35] Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[36] Xiangrong Yin, Bing Li, and R. Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99:1733–1757, 09 2008.

# 8 Appendix

*Justification for* (8)*:* Theorem 3.1 of [21] and the fact that $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, where $\mathcal{B}(\mathbb{R}^p)$ denotes the Borel sets on $\mathbb{R}^p$, is a Polish space guarantee the existence of the regular conditional probability of $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ [see also [16]]. Further, the measure is concentrated on the affine subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\} \subset \mathbb{R}^p$ and is given by (8) by Definition 8.38 and Theorem 8.39 of [20] and the orthogonal decomposition (5).

*Proof of* (9)*:* Since $\mathbf{X}$ and $\epsilon$ in (1) are assumed to be independent, $\mathbb{V}\text{ar}(Y \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \mathbb{V}\text{ar}(g(\mathbf{B}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) + \mathbb{V}\text{ar}(\epsilon)$. Using (8) and $\mathbb{V}\text{ar}(Y \mid Z) = \mathbb{E}(Y^2 \mid Z) - \mathbb{E}(Y \mid Z)^2$, we obtain (9).

We let $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) = g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})$. The parameter integral (11) is well defined and continuous if (1) $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \cdot)$ is integrable for all $\mathbf{V} \in \mathcal{S}(p, q), \mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$, (2) $\tilde{g}(\cdot, \cdot, \mathbf{r})$ is continuous for all $\mathbf{r}$, and (3) there exists an integrable dominating function of $\tilde{g}$ that does not depend on $\mathbf{V}$ and $\mathbf{s}_0$ [see [19, p. 101]].

Furthermore $t^{(l)}(\mathbf{V}, \mathbf{s}_0) = \int_{\mathcal{K}} \tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) d\mathbf{r}$ for some compact set $\mathcal{K}$, since $\text{supp}(f_{\mathbf{X}})$ is compact due to (A.4). The function $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r})$ is continuous in all inputs by the continuity of $g$ and $f_{\mathbf{X}}$ by (A.2), and therefore it attains a maximum. In consequence, all three conditions are satisfied so that $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$ is well defined and continuous.

Next $\mu_l(\mathbf{V}, \mathbf{s}_0) = t^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ is continuous since $t^{(0)}(\mathbf{V}, \mathbf{s}_0) > 0$ for all $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ by the continuity of $f_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{x}} > 0$. Then, $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (9) is continuous, which results in $L(\mathbf{V})$ also being well defined and continuous by virtue of it being a parameter integral following the same arguments as above. □

Next we establish the consistency of the conditional variance estimator. The uniform convergence in probability of the sample objective function in (19) is a sufficient condition for obtaining the consistency of $\widehat{\mathbf{V}}_q = \text{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L_n(\mathbf{V})$, as uniform convergence in probability of a random function implies convergence in probability of the minimizer of $L_n(\mathbf{V})$ to the minimizer of the limit function. Let

$$t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = \frac{1}{n h_n^{(p-q)/2}} \sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) Y_i^l \tag{37}$$

be the sample version of (11) for $l = 0, 1, 2$. The summands of $\tilde{L}_n$ in (18) can be expressed as

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)}, \tag{38}$$

Before we start with the proof a few auxiliary lemmas are shown.

**Lemma 8.** *Assume (A.4) and (K.1) hold. Let* $Z_n(\mathbf{V}, \mathbf{s}_0) = \left(\sum_i g(\mathbf{X}_i)^l K(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)\right) / (n h_n^{(p-q)/2})$ *for a continuous function* $g$*. Then,*

$$\mathbb{E}\left(Z_n(\mathbf{V}, \mathbf{s}_0)\right) = \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2) \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^q} \tilde{g}(\mathbf{r}_1, h_n^{1/2}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

where $\tilde{g}(\mathbf{r}_1, \mathbf{r}_2) = g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)$, $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$ in (5).

*Proof of Lemma 8.* By (5), $\|\mathbf{P}_{\mathbf{U}}(\mathbf{x} - \mathbf{s}_0)\|^2 = \|\mathbf{U}\mathbf{r}_2\|^2 = \|\mathbf{r}_2\|^2$. Further

$$\begin{aligned}
\mathbb{E}\left(Z_n(\mathbf{V}, \mathbf{s}_0)\right) &= \frac{1}{h_n^{(p-q)/2}} \int_{\text{supp}(f_{\mathbf{X}})} g(\mathbf{x})^l K(\|\mathbf{P}_U(\mathbf{x} - \mathbf{s}_0)/h_n^{1/2}\|^2) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \frac{1}{h_n^{(p-q)/2}} \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)^l K(\|\mathbf{r}_2/h_n^{1/2}\|^2) \times \\
&\qquad f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\
&= \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2) \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2)^l \times \\
&\qquad f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2
\end{aligned}$$

where the substitution $\tilde{\mathbf{r}}_2 = \mathbf{r}_2/h_n^{1/2}$, $d\mathbf{r}_2 = h_n^{(p-q)/2} d\tilde{\mathbf{r}}_2$ was used to obtain the last equality. □

**Lemma 9.** *Assume (A.1), (A.2), (A.3), (A.4), (H.1) and (K.1) hold. For all $\delta > 0$ there exist an $n^\star$ and finite constants $\tilde{b}^{u,m}$ for $u \in \{0,1,2,3,4\}$ and $m \in \{1,2\}$ such that*

$$\frac{(\tilde{b}^{2l,2} - \delta)}{nh_n^{(p-q)/2}} - \frac{(\tilde{b}^{l,1})^2 + \delta}{n} \leq \mathbb{V}\mathrm{ar}(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)) \leq \frac{(\tilde{b}^{2l,2} + \delta)}{nh_n^{(p-q)/2}} - \frac{(\tilde{b}^{l,1})^2 - \delta}{n}$$

for $n > n^\star$ and $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$, $l = 0, 1, 2$, in (37).

*Proof of Lemma 9.* From (1) and the binomial formula, $Y_i^l = (g_i + \epsilon_i)^l = \sum_{u=0}^l \binom{l}{u} g_i^{l-u} \epsilon_i^u$ with $g_i = g(\mathbf{B}^T \mathbf{X}_i)$. For $m \in \{1,2\}$ and $l \in \{0,1,...,4\}$, using the independence of $\mathbf{X}_i$ from $\epsilon_i$, we obtain

$$\mathbb{E}\left(Y_i^l K^m(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)\right) = \sum_{u=0}^{l} \binom{l}{u} \mathbb{E}\left(g_i^{l-u} K^m(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)\right) \mathbb{E}\left(\epsilon_i^u\right). \tag{39}$$

Setting $Z_n(\mathbf{V}, \mathbf{s}_0) = 1/(nh_n^{(p-q)/2}) \sum_i g(\mathbf{X}_i)^{l-u} \tilde{K}(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)$ in Lemma 8, where $\tilde{K}(z) = K^m(z)$ fulfills (K.1) for $m = 1, 2$, we obtain $\mathbb{E}\left(g_i^{l-u} K^m(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)\right) = h_n^{(p-q)/2} \mathbb{E}(Z_n)$. That is, if a kernel satisfies (K.1) its square also satisfies (K.1). Since the integrals are over compact sets by (A.4), by the dominated convergence theorem and Lemma 8, it holds

$$\mathbb{E}(Z_n) = \int_{\mathrm{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} \tilde{K}(\|r_2\|^2) \int_{\mathrm{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} \tilde{g}(\mathbf{r}_1, h_n^{1/2} \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 = b_n^{l-u,m} \tag{40}$$

$$\xrightarrow[n \to \infty]{} b^{l-u,m} = \int_{\mathrm{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} \tilde{K}(\|\mathbf{r}_2\|^2) d\mathbf{r}_2 \int_{\mathrm{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} \tilde{g}(\mathbf{r}_1, 0) dr_1 \tag{41}$$

using that $\tilde{g}(\mathbf{r}_1, \mathbf{r}_2) = g(\mathbf{B}^T \mathbf{s}_0 + \mathbf{B}^T \mathbf{V} \mathbf{r}_1 + \mathbf{B}^T \mathbf{U} \mathbf{r}_2)^{l-u} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)$ is continuous by (A.2) and $h_n \to 0$ by assumption (H.1).

Assumption (A.3) implies $\mathbb{E}(\epsilon_i^4) < \infty$ for $i = 1, \ldots, n$. From (40) and (39), we obtain

$$\mathbb{E}\left(Y_i^l K^m(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)/h_n^{(p-q)/2}\right) = \sum_{u=0}^{l} \binom{l}{u} b_n^{l-u,m} \mathbb{E}\left(\epsilon_i^u\right) = \tilde{b}_n^{l,m} \tag{42}$$

$$\xrightarrow[n \to \infty]{} \sum_{u=0}^{l} \binom{l}{u} b^{l-k,m} \mathbb{E}\left(\epsilon_i^u\right) = \tilde{b}^{l,m} < \infty$$

By (42), we have for $l = 0, 1, 2$

$$\mathbb{V}\mathrm{ar}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) = \frac{1}{nh_n^{p-q}} \mathbb{V}\mathrm{ar}\left(Y_1^l K(d_1(\mathbf{V}, \mathbf{s}_0)/h_n)\right) = \frac{\tilde{b}_n^{2l,2}}{nh_n^{(p-q)/2}} - \frac{(\tilde{b}_n^{l,1})^2}{n}$$

since $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$ are independent draws from the joint distribution of $(Y, \mathbf{X})$. This completes the proof since $\tilde{b}_n^{u,m} \to \tilde{b}^{u,m} < \infty$ for $u \in \{0,1,...,4\}$ and $m \in \{1,2\}$. □

Next we show that $d_i(\mathbf{V}, \mathbf{s}_0)$ in (16) is Lipschitz in its inputs under assumption (A.4) in Lemma 10.

**Lemma 10.** *Under assumption (A.4) there exists a constant $0 < C_2 < \infty$ such that for all $\delta > 0$ and $\mathbf{V}, \mathbf{V}_j \in \mathcal{S}(p,q)$ with $\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| < \delta$ and for all $\mathbf{s}_0, \mathbf{s}_j \in$ supp$(f_{\mathbf{X}}) \subset \mathbb{R}^p$ with $\|\mathbf{s}_0 - \mathbf{s}_j\| < \delta$*

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2 \delta$$

*for $d_i(\mathbf{V}, \mathbf{s}_0)$ given by* (16)

*Proof of Lemma 10.*

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq \left|\|\mathbf{X}_i - \mathbf{s}_0\|^2 - \|\mathbf{X}_i - \mathbf{s}_j\|^2\right| +$$
$$\left|\langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{P}_{\mathbf{V}}(\mathbf{X}_i - \mathbf{s}_0)\rangle - \langle \mathbf{X}_i - \mathbf{s}_j, \mathbf{P}_{\mathbf{V}_j}(\mathbf{X}_i - \mathbf{s}_j)\rangle\right| = I_1 + I_2 \tag{43}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product on $\mathbb{R}^p$. For the first term on the right hand side of (43)

$$I_1 = \left|\|\mathbf{X}_i - \mathbf{s}_0\|^2 - \|\mathbf{X}_i - \mathbf{s}_j\|^2\right| \leq 2\left|\langle \mathbf{X}_i, \mathbf{s}_0 - \mathbf{s}_j\rangle\right| + \left|\|\mathbf{s}_0\|^2 - \|\mathbf{s}_j\|^2\right|$$
$$\leq 2\|\mathbf{X}_i\|\|\mathbf{s}_0 - \mathbf{s}_j\| + 2C_1\|\mathbf{s}_0 - \mathbf{s}_j\| \leq 2C_1\delta + 2C_1\delta = 4C_1\delta$$

by Cauchy-Schwartz and the reverse triangular inequality (i.e. $\left|\|\mathbf{s}_0\|^2 - \|\mathbf{s}_j\|^2\right| = |\|\mathbf{s}_0\| - \|\mathbf{s}_j\|| \left(\|\mathbf{s}_0\| + \|\mathbf{s}_j\|\right) \leq \|\mathbf{s}_0 - \mathbf{s}_j\| 2C_1$) and $\|\mathbf{X}_i\| \leq \sup_{z \in \text{supp}(f_{\mathbf{X}})} \|z\| = C_1 < \infty$ with probability 1 due to (A.4). The second term in (43) satisfies

$$\begin{aligned} I_2 &\leq \left|\langle \mathbf{X}_i, (\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j})\mathbf{X}_i \rangle\right| + 2\left|\langle \mathbf{X}_i, \mathbf{P}_{\mathbf{V}}\mathbf{s}_0 - \mathbf{P}_{\mathbf{V}_j}\mathbf{s}_j \rangle\right| + \left|\langle \mathbf{s}_0, \mathbf{P}_{\mathbf{V}}\mathbf{s}_0 \rangle - \langle \mathbf{s}_j, \mathbf{P}_{\mathbf{V}_j}\mathbf{s}_j \rangle\right| \\ &\leq \|\mathbf{X}_i\|^2 \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| + 2\|\mathbf{X}_i\| \left\|\mathbf{P}_{\mathbf{V}}(\mathbf{s}_0 - \mathbf{s}_j) + (\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j})\mathbf{s}_j\right\| + |\langle \mathbf{s}_0 - \mathbf{s}_j, \mathbf{P}_{\mathbf{V}}\mathbf{s}_0 \rangle| + \\ &\left|\langle \mathbf{s}_j, \mathbf{P}_{\mathbf{V}}\mathbf{s}_0 - \mathbf{P}_{\mathbf{V}_j}\mathbf{s}_j \rangle\right| \leq C_1^2\delta + 2C_1(\delta + C_1\delta) + C_1\delta + C_1(\delta + C_1\delta) = 4C_1\delta + 4C_1^2\delta \end{aligned}$$

Collecting all constants into $C_2$ (i.e. $C_2 = 8C_1 + 4C_1^2$) yields the result. □

The proofs of Theorems 3 and 11 require the **Bernstein inequality** [3]: Let $Z_1, Z_2, ...$ be an independent sequence of bounded random variables $|Z_i| \leq b$. Let $S_n = \sum_{i=1}^n Z_i$, $E_n = \mathbb{E}(S_n)$ and $V_n = \mathbb{V}\text{ar}(S_n)$. Then,

$$P(|S_n - E_n| > t) < 2\exp\left(-\frac{t^2/2}{V_n + bt/3}\right) \tag{44}$$

Furthermore the proof of Theorem 11 requires assumption (K.2), which obtains

$$|K(u) - K(u')| \leq K^*(u')\delta \tag{45}$$

for all $u, u'$ with $|u - u'| < \delta \leq L_2$ and $K^*(\cdot)$ is a bounded and integrable kernel function [see [18]]. Specifically, if condition (1) of (K.2) holds, then $K^*(u) = L_1 1_{\{|u| \leq 2L_2\}}$. If (2) holds, then $K^*(u) = L_1 1_{\{|u| \leq 2L_2\}} + 1_{\{|u| > 2L_2\}}|u - L_2|^{-\nu}$.

Let $A = \mathcal{S}(p,q) \times \text{supp}(f_{\mathbf{X}})$ and by a slight abuse of notation, we generically denote constants by $C$. In Theorems 11 and 12 we show that the variance and bias terms of (37) vanish uniformly in probability, respectively.

**Theorem 11.** *Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$ and $a_n/h_n^{(p-q)/2} = O(1)$,*

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| = O_P(a_n) \quad \textit{for} \quad l = 0, 1, 2 \tag{46}$$

**Remark.** *If we assume $|Y| < M_2 < \infty$ almost surely, the requirement $a_n/h_n^{(p-q)/2} = O(1)$ for the bandwidth can be dropped and the truncation step of the proof of Theorem 11 can be skipped.*

*Proof of Theorem 11.* The proof is organized in 3 steps: a truncation step, a discretization step by covering $A = \mathcal{S}(p,q) \times \text{supp}(f_{\mathbf{X}})$, and application of Bernstein's inequality (44).

We let $\tau_n = a_n^{-1}$ and truncate $Y_i^l$ by $\tau_n$ as follows. We let

$$t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) = (1/nh_n^{(p-q)/2}) \sum_i K(\|\mathbf{P}_{\mathbf{U}}(\mathbf{X}_i - \mathbf{s}_0)\|^2/h_n) Y_i^l 1_{\{|Y_i|^l \leq \tau_n\}} \tag{47}$$

be the truncated version of (37) and $\tilde{R}_n^{(l)} = (1/nh_n^{(p-q)/2}) \sum_i |Y_i|^l 1_{\{|Y_i|^l > \tau_n\}}$ be the remainder of (37). Therefore $R_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) \leq M_1 \tilde{R}_n^{(l)}$ due to (K.1) and

$$\begin{aligned} \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| &\leq M_1(\tilde{R}_n^{(l)} + \mathbb{E}\tilde{R}_n^{(l)}) \\ &+ \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left|t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \end{aligned} \tag{48}$$

By Cauchy-Schwartz and the Markov inequality, $\mathbb{P}(|Z| > t) = \mathbb{P}(Z^4 > t^4) \leq \mathbb{E}(Z^4)/t^4$, we obtain

$$\begin{aligned} \mathbb{E}\tilde{R}_n^{(l)} &= \frac{1}{h_n^{(p-q)/2}} \mathbb{E}\left(|Y_i|^l 1_{\{|Y_i|^l > \tau_n\}}\right) \leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|Y_i|^{2l})} \sqrt{\mathbb{P}(|Y_i|^l > \tau_n)} \\ &\leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|Y_i|^{2l})} \left(\frac{\mathbb{E}(|Y_i|^{4l})}{a_n^{-4}}\right)^{1/2} = o(a_n) \end{aligned} \tag{49}$$

where the last equality uses the assumption $a_n/h_n^{(p-q)/2} = O(1)$ and the expectations are finite due to (A.3) for $l = 0, 1, 2$. Obviously, no truncation is needed for $l = 0$.

Therefore the first two terms of the right hand side of (48) converge to 0 with rate $a_n$ by (49) and Markov's inequality. From now to the end of the proof $Y_i$ will denote the truncated version $Y_i 1_{\{|Y_i| \leq \tau_n\}}$ and we do not distinguish the truncated from the untruncated $t_n(\mathbf{V}, \mathbf{s}_0)$ since this truncation results in an error of magnitude $a_n$.

For the discretization step we cover the compact set $A = \mathcal{S}(p,q) \times \text{supp}(f_{\mathbf{X}})$ by finitely many balls, which is possible by (A.4) and the compactness of $\mathcal{S}(p,q)$. Let $\delta_n = a_n h_n$ and $A_j = \{\mathbf{V} : \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \leq \delta_n\} \times \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}_j\| \leq \delta_n\}$ be a cover of $A$ with ball centers $\mathbf{V}_j \times \mathbf{s}_j$. Then, $A \subset \bigcup_{j=1}^N A_j$ and the number of balls can be bounded by $N \leq C\, \delta_n^{-d} \delta_n^{-p}$ for some constant $C \in (0, \infty)$, where $d = \dim(\mathcal{S}(p,q)) = pq - q(q+1)/2$. Let $\mathbf{V} \times \mathbf{s}_0 \in A_j$. Then by Lemma 10 there exists $0 < C_2 < \infty$, such that

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2 \delta_n \tag{50}$$

holds for $d_i$ in (16). Under (K.2), which implies (45), inequality (50) yields

$$\left| K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) - K\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right)\right| \leq K^*\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right) C_2 a_n \tag{51}$$

for $\mathbf{V} \times \mathbf{s}_0 \in A_j$ and $K^*(\cdot)$ an integrable and bounded function.

Define $r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) = (1/nh_n^{(p-q)/2}) \sum_{i=1}^n K^*(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)|Y_i|^l$. For notational convenience we drop the dependence on $l$ and $j$ in the following and observe that (51) yields

$$|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2 a_n r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) \tag{52}$$

Since $K^*$ fulfills (K.1) except for continuity, an analogous argument as in the proof of Lemma 8 yields that $\mathbb{E}\left(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right) < \infty$ by $(A.3)$. By subtracting and adding $t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)$, $\mathbb{E}(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$, the triangular inequality, (52) and integrability of $r_n^l$, we obtain

$$\begin{aligned}
&\left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \leq \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right| + \left|\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \\
&+ \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \leq C_2 a_n \left(|r_n| + |\mathbb{E}\left(r_n\right)|\right) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq C_2 a_n (|r_n - \mathbb{E}(r_n)| + 2|\mathbb{E}(r_n)|) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq 2C_3 a_n + |r_n - \mathbb{E}(r_n)| + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right|
\end{aligned} \tag{53}$$

for any $C_3 > C_2 \mathbb{E}(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$ and $n$ such that $a_n \leq 1$, since $a_n^2 = o(1)$. Summarizing there exists $0 < C_3 < \infty$ such that (53) holds.

Then using $\sup_{x \in A} f(x) = \max_{1 \leq j \leq N} \sup_{x \in A_j} f(x) \leq \sum_{j=1}^N \sup_{x \in A_j} f(x)$ for any partition of $A$ and continuous function $f$, subadditivty of the probability for the first inequality and (53) for the third inequality below, it holds

$$\begin{aligned}
&\mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n) \\
&\leq \sum_{j=1}^N \mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n) \\
&\leq N \max_{1 \leq j \leq N} \mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n) \\
&\leq N \left(\max_{1 \leq j \leq N} \mathbb{P}(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n) + \max_{1 \leq j \leq N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\right) \leq \\
&C\, \delta^{-(d+p)} \left(\max_{1 \leq j \leq N} \mathbb{P}(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n) + \max_{1 \leq j \leq N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\right)
\end{aligned} \tag{54}$$

where the last inequality is due to $N \leq C\, \delta_n^{-d} \delta_n^{-p}$ for a cover of $A$.

Finally, we bound the first and second term in the last line of (54) by the Bernstein inequality (44). For the first term in the last line of (54), let $Z_i = Y_i^l K(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)$ and $S_n = \sum_i Z_i = nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)$, then the $Z_i$ are independent, $|Z_i| \leq b = M_1 \tau_n = M_1/a_n$ by (K.1) and the truncation step. For $V_n = \mathbb{V}\text{ar}(S_n)$, Lemma 9 yields

$$nh_n^{(p-q)/2}\left(\tilde{b}^{2l,2} - \delta - h_n^{(p-q)/2}\left((\tilde{b}^{l,1})^2 + \delta\right)\right) \leq V_n \leq nh_n^{(p-q)/2}\left(\tilde{b}^{2l,2} + \delta - h_n^{(p-q)/2}\left((\tilde{b}^{l,1})^2 - \delta\right)\right)$$

for $n$ sufficiently large. We write $nh_n^{(p-q)/2}C \geq V_n$ with $C = \tilde{b}^{2l,2} + \delta$, and set $t = C_3 a_n nh_n^{(p-q)/2}$. The Bernstein inequality (44) yields

$$\mathbb{P}\left(\left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| > C_3 a_n\right) < 2\exp\left(\frac{-t^2/2}{V_n + bt/3}\right) \leq$$
$$2\exp\left(-\frac{(1/2)C_3^2 a_n^2 n^2 h_n^{(p-q)}}{nh_n^{(p-q)/2}C + (1/3)M_1\tau_n C_3 a_n nh_n^{(p-q)/2})}\right) \leq 2\exp\left(-\frac{(1/2)C_3\log(n)}{C/C_3 + (M_1/3)}\right) = 2n^{-\gamma(C_3)}$$

where $a_n^2 = \log(n)/(nh_n^{(p-q)/2})$ and define $\gamma(C_3) = \frac{(1/2)C_3}{C/C_3+(M_1/3)}$, which is an increasing function that can be made arbitrarily large by increasing $C_3$.

For the second term in the last line of (54), set $Z_i = Y_i^l K^*(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)$ in the Bernstein inequality (44) and proceed analogously to obtain

$$\mathbb{P}\left(\left|r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| > C_3 a_n\right) < 2n^{-\frac{(1/2)C_3}{C/C_3+(1/3)M_2}} = 2n^{-\gamma(C_3)}$$

By (H.1), $h_n^{(p-q)/2} \leq 1$ for $n$ large, so that $\delta_n^{-1} = (a_n h_n)^{-1} \leq n^{1/2}h_n^{-1}h_n^{(p-q)/4} \leq n^{5/2}$. Further (H.2) implies $1/(nh_n^{(p-q)/2}) \leq 1$ for $n$ large, therefore $h_n^{-1} \leq n^{2/(p-q)} \leq n^2$ since $p - q \geq 1$. Therefore, (54) is smaller than $4\,C\,\delta_n^{-(d+p)}n^{-\gamma(C_3)} \leq 4Cn^{5(d+p)/2-\gamma(C_3)}$. For $C_3$ large enough, we have $5(d+p)/2 - \gamma(C_3) < 0$ and $n^{5(d+p)/2-\gamma(C_3)} \to 0$. This completes the proof. □

**Theorem 12.** *Under (A.1), (A.2) and (A.4), (H.1), (K.1), and* $\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)d\mathbf{r}_2 = 1$,

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|t^{(l)}(\mathbf{V}, \mathbf{s}_0) + 1_{\{l=2\}}\eta^2 t^{(0)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| = O(h_n), \quad l = 0, 1, 2 \tag{55}$$

*Proof of Theorem 12.* Let $\tilde{g}(\mathbf{r}_1, \mathbf{r}_2) = g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1 + \mathbf{B}^T\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)$ where $\mathbf{r}_1, \mathbf{r}_2$ satisfy the orthogonal decomposition (5). Then

$$\mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) = \int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\tilde{g}(\mathbf{r}_1, {h_n}^{1/2}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 + 1_{\{l=2\}}\eta^2\mathbb{E}\left(t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)\right) \tag{56}$$

holds by Lemma 8 for $l = 0, 1$. For $l = 2$, $Y_i^2 = g_i^2 + 2g_i\epsilon_i + \epsilon_i^2$ with $g_i = g(\mathbf{B}^T\mathbf{X}_i)$ and can be handled as in the case of $l = 0, 1$.

Plugging in (56) the second order Taylor expansion for some $\xi$ in the neighborhood of 0, $\tilde{g}(\mathbf{r}_1, {h_n}^{1/2}\mathbf{r}_2) = \tilde{g}(\mathbf{r}_1, 0) + {h_n}^{1/2}\nabla_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1, 0)^T\mathbf{r}_2 + h_n\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1, \xi)\mathbf{r}_2$, yields

$$\mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) = \int_{\mathbb{R}^q}\tilde{g}(\mathbf{r}_1, 0)d\mathbf{r}_1 + \sqrt{h_n}\left(\int_{\mathbb{R}^q}\nabla_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1, 0)d\mathbf{r}_1\right)^T\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\mathbf{r}_2 d\mathbf{r}_2 +$$
$$h_n\frac{1}{2}\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1, \xi)\mathbf{r}_2 d\mathbf{r}_1 d\mathbf{r}_2 = t^{(l)}(\mathbf{V}, \mathbf{s}_0) + h_n\frac{1}{2}R(\mathbf{V}, \mathbf{s}_0)$$

since $\int_{\mathbb{R}^q}\tilde{g}(\mathbf{r}_1, 0)d\mathbf{r}_1 = t^{(l)}(\mathbf{V}, \mathbf{s}_0)$ and $\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\mathbf{r}_2 d\mathbf{r}_2 = 0 \in \mathbb{R}^{p-q}$ due to $K(\|\cdot\|^2)$ being even. Let $R(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1, \xi)\mathbf{r}_2 d\mathbf{r}_1 d\mathbf{r}_2$. By (A.4) and (A.2) it holds $|\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1, \xi)\mathbf{r}_2| \leq C\|\mathbf{r}_2\|^2$ for $C = \sup_{\mathbf{x},\mathbf{y}}\|\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{x}, \mathbf{y})\| < \infty$, since a continuous function over a compact set is bounded. Then, $R(\mathbf{V}, \mathbf{s}_0) \leq CC_4\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\|\mathbf{r}_2\|^2 d\mathbf{r}_2 < \infty$ for some $C_4 > 0$ since the integral over $\mathbf{r}_1$ is over a compact set by (A.4). □

Lemma 13 follows directly from Theorems 11 and 12 and the triangle inequality.

**Lemma 13.** *Suppose (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1) hold. If* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$, *and* $a_n/h_n^{(p-q)/2} = O(1)$, *then for* $l = 0, 1, 2$

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|t^{(l)}(\mathbf{V}, \mathbf{s}_0) + 1_{\{l=2\}}\eta^2 t^{(0)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right| = O_P(a_n + h_n)$$

Combining the results of Theorems 11 and 12 and Lemma 13 obtains Theorem 14.

**Theorem 14.** *Suppose (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1) hold. Let* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$, $a_n/h_n^{(p-q)/2} = O(1)$, $\delta_n = \inf_{\mathbf{V}\times\mathbf{s}_0\in A_n} t^{(0)}(\mathbf{V},\mathbf{s}_0)$, *where* $t^{(0)}(\mathbf{V},\mathbf{s}_0)$ *is defined in* (11), *andand* $A_n = \mathcal{S}(p,q)\times\{\mathbf{x}\in supp(f_{\mathbf{X}}) : |\mathbf{x}-\partial supp(f_{\mathbf{X}})|\geq b_n\}$, *where* $\partial C$ *denotes the boundary of the set* $C$ *and* $|\mathbf{x}-C| = \inf_{\mathbf{r}\in C}|\mathbf{x}-\mathbf{r}|$, *for a sequence* $b_n\to 0$ *so that* $\delta_n^{-1}(a_n+h_n)\to 0$ *for any bandwidth* $h_n$ *that satisfies the assumptions. Then,*

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|\bar{y}_l(\mathbf{V},\mathbf{s}_0)-\mu_l(\mathbf{V},\mathbf{s}_0)-1_{\{l=2\}}\eta^2 t^{(0)}(\mathbf{V},\mathbf{s}_0)\right| = O_P(\delta_n^{-1}(a_n+h_n)),\quad l=0,1,2$$

*and*

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|\tilde{L}_n(\mathbf{V},\mathbf{s}_0)-\tilde{L}(\mathbf{V},\mathbf{s}_0)\right| = O_P(\delta_n^{-1}(a_n+h_n)) \tag{57}$$

*where* $\bar{y}_l(\mathbf{V},\mathbf{s}_0)$, $\mu_l(\mathbf{V},\mathbf{s}_0)$, $\tilde{L}_n(\mathbf{V},\mathbf{s}_0)$ *and* $\tilde{L}(\mathbf{V},\mathbf{s}_0)$ *are defined in* (38), (10), (18) *and* (9), *respectively.*

*Proof of Theorem 14.*

$$\bar{y}_l(\mathbf{V},\mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)}{t_n^{(0)}(\mathbf{V},\mathbf{s}_0)} = \frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)/t^{(0)}(\mathbf{V},\mathbf{s}_0)}{t_n^{(0)}(\mathbf{V},\mathbf{s}_0)/t^{(0)}(\mathbf{V},\mathbf{s}_0)}$$

We consider the numerator and enumerator separately. By Lemma 13

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A_n}\left|\frac{t_n^{(0)}(\mathbf{V},\mathbf{s}_0)}{t^{(0)}(\mathbf{V},\mathbf{s}_0)}-1\right| \leq \frac{\sup_A|t_n^{(0)}(\mathbf{V},\mathbf{s}_0)-t^{(0)}(\mathbf{V},\mathbf{s}_0)|}{\inf_{A_n} t^{(0)}(\mathbf{V},\mathbf{s}_0)} = O_P(\delta_n^{-1}(a_n+h_n))$$

Next

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A_n}\left|\frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)}{t^{(0)}(\mathbf{V},\mathbf{s}_0)}-\mu_l(\mathbf{V},\mathbf{s}_0)\right| \leq \frac{\sup_A|t_n^{(l)}(\mathbf{V},\mathbf{s}_0)-t^{(l)}(\mathbf{V},\mathbf{s}_0)|}{\inf_{A_n} t^{(0)}(\mathbf{V},\mathbf{s}_0)} = O_P(\delta_n^{-1}(a_n+h_n)).$$

Therefore by $A_n\uparrow A = \mathcal{S}(p,q)\times \text{supp}(f_{\mathbf{X}})$ we get

$$\lim_{n\to\infty}\sup_{\mathbf{V}\times\mathbf{s}_0\in A_n}\left|\frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)}{t^{(0)}(\mathbf{V},\mathbf{s}_0)}-\mu_l(\mathbf{V},\mathbf{s}_0)\right| = \lim_{n\to\infty}\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|\frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)}{t^{(0)}(\mathbf{V},\mathbf{s}_0)}-\mu_l(\mathbf{V},\mathbf{s}_0)\right|$$

and in total we obtain

$$\bar{y}_l(\mathbf{V},\mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)/t^{(0)}(\mathbf{V},\mathbf{s}_0)}{t_n^{(0)}(\mathbf{V},\mathbf{s}_0)/t^{(0)}(\mathbf{V},\mathbf{s}_0)} = \frac{\mu_l+O_P(\delta_n^{-1}(a_n+h_n))}{1+O_P(\delta_n^{-1}(a_n+h_n))} = \mu_l+O_P(\delta_n^{-1}(a_n+h_n)).$$

For $l=2$, $Y_i^2 = g(\mathbf{B}^T\mathbf{X}_i)^2+2g(\mathbf{B}^T\mathbf{X}_i)\epsilon_i+\epsilon_i^2$, and (57) follows from (9). □

**Lemma 15.** *Under (A.1), (A.2), (A.4), there exists* $0<C_5<\infty$ *such that*

$$|\mu_l(\mathbf{V},\mathbf{s}_0)-\mu_l(\mathbf{V}_j,\mathbf{s}_0)| \leq C_5\|\mathbf{P}_{\mathbf{V}}-\mathbf{P}_{\mathbf{V}_j}\| \tag{58}$$

*for all* $\mathbf{s}_0\in supp(f_{\mathbf{X}})$

*Proof.* From the representation $\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)$ in (14) instead of $t^{(l)}(\mathbf{V},\mathbf{s}_0)$, we consider $\mu_l(\mathbf{V},\mathbf{s}_0)=\mu_l(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)$ as a function on the Grassmann manifold. Then,

$$\begin{aligned}
\left|\mu_l(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)-\mu_l(\mathbf{P}_{\mathbf{V}_j},\mathbf{s}_0)\right| &= \left|\frac{\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)}-\frac{\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j},\mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j},\mathbf{s}_0)}\right| \\
&\leq \frac{\sup|\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)|}{(\inf\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0))^2}\left|\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)-\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j},\mathbf{s}_0)\right| \\
&\quad+\frac{\sup\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)}{(\inf\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0))^2}\left|\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)-\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j},\mathbf{s}_0)\right|
\end{aligned} \tag{59}$$

with $\sup_{\mathbf{P}_{\mathbf{V}}\in Gr(p,q)}\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)\in(0,\infty)$ and $\inf_{\mathbf{P}_{\mathbf{V}}\in Gr(p,q)}\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}},\mathbf{s}_0)\in(0,\infty)$ since $\tilde{t}^{(l)}$ is continuous, $\mathbf{\Sigma}_{\mathbf{x}}>0$ and $\mathbf{s}_0\in\text{supp}(f_{\mathbf{X}})$.

By (A.2), $\tilde{g}(\mathbf{x}) = g(\mathbf{B}^T\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ is twice continuous differentiable and therefore Lipschitz continuous on compact sets. We denote its Lipschitz constant by $L < \infty$. Therefore,

$$\left|\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0) - \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)\right| \leq \int_{\text{supp}(f_{\mathbf{X}})} \left|\tilde{g}(\mathbf{s}_0 + \mathbf{P}_{\mathbf{V}}\mathbf{r}) - \tilde{g}(\mathbf{s}_0 + \mathbf{P}_{\mathbf{V}_j}\mathbf{r})\right| d\mathbf{r}$$

$$\leq L \int_{\text{supp}(f_{\mathbf{X}})} \|(\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j})\mathbf{r}\| d\mathbf{r} \leq L \left(\int_{\text{supp}(f_{\mathbf{X}})} \|\mathbf{r}\| dr\right) \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{60}$$

where the last inequality is due to the sub-multiplicativity of the Frobenius norm and the integral being finite by (A.4). Plugging (60) in (59) and collecting all constants into $C_5$ yields (58). □

*Proof of Theorem 3.* By (19) and (7),

$$|L_n(\mathbf{V}) - L(\mathbf{V})| \leq \left|\frac{1}{n}\sum_i \left(\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)\right)\right| + \left|\frac{1}{n}\sum_i \left(\tilde{L}(\mathbf{V}, \mathbf{X}_i) - \mathbb{E}(\tilde{L}(\mathbf{V}, \mathbf{X}))\right)\right| \tag{61}$$

The first term on the right hand side of (61) goes to 0 in probability uniformly in $\mathbf{V}$ by Theorem 14,

$$\left|\frac{1}{n}\sum_i \tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)\right| \leq \sup_{\mathbf{V}\times\mathbf{s}_0 \in A} \left|\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) - \tilde{L}(\mathbf{V}, \mathbf{s}_0)\right| = O_P(\delta_n^{-1}(a_n + h_n)) \tag{62}$$

The second term in (61) converges to 0 almost surely for all $\mathbf{V} \in \mathcal{S}(p, q)$ by the strong law of large numbers. In order to show uniform convergence the same technique as in the proof of Theorem 11 is used. Let $B_j = \{\mathbf{V} \in \mathcal{S}(p, q) : \|\mathbf{V}\mathbf{V}^T - \mathbf{V}_j\mathbf{V}_j^T\| \leq \tilde{a}_n\}$ be a cover of $\mathcal{S}(p, q) \subset \bigcup_{j=1}^N B_j$ with $N \leq C\,\tilde{a}_n^{-d} = C\,(n/\log(n))^{d/2} \leq C\,n^{d/2}$, where $d = \dim(\mathcal{S}(p, q))$ is defined in the proof of Theorem 11. By Lemma 15,

$$|\mu_l(\mathbf{V}, \mathbf{X}_i) - \mu_l(\mathbf{V}_j, \mathbf{X}_i)| \leq C_5\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{63}$$

Let $G_n(\mathbf{V}) = \sum_i \tilde{L}(\mathbf{V}, \mathbf{X}_i)/n$ with $\mathbb{E}(G_n(V)) = L(\mathbf{V})$. Using (63) and following the same steps as in the proof of Theorem 11 we obtain

$$\begin{aligned} |G_n(\mathbf{V}) - L(\mathbf{V})| &\leq |G_n(\mathbf{V}) - G_n(\mathbf{V}_j)| + |G_n(\mathbf{V}_j) - L(\mathbf{V}_j)| + |L(\mathbf{V}) - L(\mathbf{V}_j)| \\ &\leq 2C_6\tilde{a}_n + |G_n(\mathbf{V}_j) - L(\mathbf{V}_j)| \end{aligned} \tag{64}$$

for $\mathbf{V} \in B_j$ and some $C_6 > C_5$. Inequality (64) leads to

$$\begin{aligned} \mathbb{P}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |G_n(\mathbf{V}) - L(\mathbf{V})| > 3C_6\tilde{a}_n\right) &\leq C\,N\,\mathbb{P}(\sup_{\mathbf{V}\in B_j} |G_n(\mathbf{V}) - L(\mathbf{V})| > 3C_6\tilde{a}_n) \\ &\leq C\,n^{d/2}\mathbb{P}(|G_n(\mathbf{V}_j) - L(\mathbf{V}_j)| > C_6\tilde{a}_n) \leq C\,n^{d/2}n^{-\gamma(C_6)} \to 0 \end{aligned} \tag{65}$$

where the last inequality in (65) is due to the Bernstein inequality (44) with $Z_i = \tilde{L}(\mathbf{V}_j, \mathbf{X}_i)$, which is bounded since $\tilde{L}(\cdot, \cdot)$ is continuous on the compact set $A$, and $\gamma(C_6)$ a monotone increasing function of $C_6$ that can be made arbitrarily large by choosing $C_6$ accordingly. Therefore, $\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_n(\mathbf{V}) - L(\mathbf{V})| \leq O_P(\delta_n^{-1}(a_n + h_n) + \tilde{a}_n)$ with $\delta_n = \inf_{\mathbf{V}\times\mathbf{s}_0\in A_n} t^{(0)}(\mathbf{V}, \mathbf{s}_0)$, where $t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ is defined in (11), and $A_n = \mathcal{S}(p, q)\times\{\mathbf{x} \in \text{supp}(f_{\mathbf{X}}) : f_{\mathbf{X}}(\mathbf{x}) \geq b_n\}$ for a sequence $b_n \to 0$ so that $\delta_n^{-1}(a_n + h_n) \to 0$ for any bandwidth $h_n$ that satisfies the assumptions, which implies (20). □

*Proof of Theorem 4.* We apply Theorem 4.1.1 of [2] to obtain consistency of the conditional variance estimator. This theorem requires three conditions that guarantee the convergence of the minimizer of a sequence of random functions $L_n(\mathbf{P}_{\mathbf{V}})$ to the minimizer of the limiting function $L(\mathbf{P}_{\mathbf{V}})$; i.e., $\mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}\}^\perp} = \text{argmin}\, L_n(\mathbf{P}_{\mathbf{V}}) \to \mathbf{P}_{\text{span}\{\mathbf{B}\}^\perp} = \text{argmin}\, L(\mathbf{P}_{\mathbf{V}})$ in probability. To apply the theorem three conditions have to be met: (1) The parameter space is compact; (2) $L_n(\mathbf{V})$ is continuous and a measurable function of the data $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$ and (3) $L_n(\mathbf{V})$ converges uniformly to $L(\mathbf{V})$ and $L(\mathbf{V})$ attains a unique global minimum at $\text{span}\{\mathbf{B}\}^\perp$.

Since $L_n(\mathbf{V})$ depends on $\mathbf{V}$ only through $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^T$, $L_n(\mathbf{V})$ can be considered as functions on the Grassmann manifold, which is compact, and the same holds true for $L(\mathbf{V})$ by (14). Further, $L_n(\mathbf{V})$ is by definition a measurable function of the data and continuous in $\mathbf{V}$ if a continuous kernel is used, such as the Gaussian. Theorem 3 obtains the uniform convergence and Theorem 1 that the minimizer is unique when $L(\mathbf{V})$ is minimized over the Grassmann manifold $G(p, q)$, since $\text{span}\{\mathbf{B}\}$ is uniquely identifiable and so is $\text{span}\{\mathbf{B}\}^\perp$ (i.e. $\|\mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}\}} - \mathbf{P}_{\text{span}\{\mathbf{B}\}}\| = \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\| = \|(\mathbf{I}_p - \mathbf{B}\mathbf{B}^T) - (\mathbf{I}_p - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T)\| = \|\mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}\}^\perp} - \mathbf{P}_{\text{span}\{\mathbf{B}\}^\perp}\|$). Thus, all three conditions are met and the result is obtained. □

*Proof of Theorem 6.* The Gaussian kernel $K$ satisfies $\partial_z K(z) = -zK(z)$. From (17) and (18) we have $\tilde{L}_n = \bar{y}_2 - \bar{y}_1^2$ where $\bar{y}_l = \sum_i w_i Y_i^l$, $l = 1, 2$. We let $K_j = K(d_j(\mathbf{V}, \mathbf{s}_0)/h_n)$, suppress the dependence on $\mathbf{V}$ and $\mathbf{s}_0$ and write $w_i = K_i / \sum_j K_j$. Then, $\nabla K_i = (-1/h_n^2) K_i d_i \nabla d_i$ and $\nabla w_i = -\left(K_i d_i \nabla d_i (\sum_j K_j) - K_i \sum_j K_j d_j \nabla d_j\right) / (h_n \sum_j K_j)^2$. Next,

$$\nabla \bar{y}_l = -\frac{1}{h_n^2} \sum_i Y_i^l \frac{\left(K_i d_i \nabla d_i - K_i (\sum_j K_j d_j \nabla d_j)\right)}{(\sum_j K_j)^2} = -\frac{1}{h_n^2} \sum_i Y_i^l w_i \left(d_i \nabla d_i - \sum_j w_j d_j \nabla d_j\right)$$

$$= -\frac{1}{h_n^2} \left(\sum_i Y_i^l w_i d_i \nabla d_i - \sum_j Y_j^l w_j \sum_i w_i d_i \nabla d_i\right) = -\frac{1}{h_n^2} \sum_i (Y_i^l - \bar{y}_l) w_i d_i \nabla d_i \tag{66}$$

Then, $\nabla \tilde{L}_n = \nabla \bar{y}_2 - 2\bar{y}_1 \nabla \bar{y}_1$, and inserting $\nabla \bar{y}_l$ from (66) yields $\nabla \tilde{L}_n = (-1/h_n^2) \sum_i (Y_i^2 - \bar{y}_2 - 2\bar{y}_1 (Y_i - \bar{y}_1)) w_i d_i \nabla d_i = (1/h_n^2)(\sum_i \left(\tilde{L}_n - (Y_i - \bar{y}_1)^2\right) w_i d_i \nabla d_i)$, since $Y_i^2 - \bar{y}_2 - 2\bar{y}_1 (Y_i - \bar{y}_1) = (Y_i - \bar{y}_1)^2 - \tilde{L}_n$. □