ORIGINAL ARTICLE

Journal Section

Conditional Variance Estimator for Sufficient Dimension Reduction

Lukas Fertl[1]* | **Efstathia Bura**[1]*

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

Correspondence
Lukas Fertl
Email: lukas.fertl@tuwien.ac.at

Present address
*TU Wien, Institute of Statistics and Mathematical Methods in Economics Faculty of Mathematics and Geoinformation
Vienna 1040, Austria, Wiednerhauptstraße 8-10

Funding information
The authors gratefully acknowledge the support of the Austrian Science Fund (FWF P 30690-N35)

Conditional Variance Estimation (CVE) is a novel sufficient dimension reduction (SDR) method for regressions satisfying $\mathbb{E}(Y|\mathbf{X}) = \mathbb{E}(Y|\mathbf{B}^\top\mathbf{X})$, where $\mathbf{B}^\top\mathbf{X}$ is a lower dimensional projection of the predictors. CVE, similarly to its main competitor, the mean average variance estimation (MAVE), is not based on inverse regression, and does not require the restrictive linearity and constant variance conditions of moment based SDR methods.CVE is data-driven and applies to additive error regressions with continuous predictors and link function. The effectiveness and accuracy of CVE compared to MAVE and other SDR techniques is demonstrated in simulation studies. CVE is shown to outperform MAVE in some model settings, while it remains on par under most others.

KEYWORDS
Regression, SDR, mean subspace, MAVE

acknowledgements

We thank Daniel Kapla for his programming assistance. He also co-authored the `CVE` `R` package that implements the proposed method.

1 | INTRODUCTION

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

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

where $\mathbf{X}$ is independent of ϵ with positive definite variance-covariance matrix, $\mathbb{V}\text{ar}(\mathbf{X}) = \mathbf{\Sigma_X}$, $\epsilon \in \mathbb{R}$ is a mean zero random variable with finite $\mathbb{V}\text{ar}(\epsilon) = \mathbb{E}\left(\epsilon^2\right) = \eta^2$, 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

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

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

Identifying the span of $\mathbf{B}$, as only the 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 $\mathbf{B}$ is semi-orthogonal; i.e., $\mathbf{B}^\top \mathbf{B} = \mathbf{I}_k$, since a change of coordinate system by an orthogonal transformation does not alter model (2).

Finding sufficient reductions of the predictors to replace them in regression and classification without loss of information is called *sufficient dimension reduction* (SDR) Cook (1998). The first split in SDR taxonomy occurs between likelihood and non-likelihood based methods. The former, which were developed more recently (Cook, 2007; Cook and Forzani, 2008, 2009; Bura and Forzani, 2015; Bura et al., 2016), assume knowledge either of the joint family of distributions of $(Y, \mathbf{X}^\top)^\top$, or the conditional family of distributions for $\mathbf{X}|Y$. The latter is the most researched branch of SDR and comprises of three classes of methods: Inverse regression based, semi-parametric and nonparametric. Reviews of the former two classes can be found in Adragni and Cook (2009); Ma and Zhu (2013); Li (2018).

In this paper we present the *conditional variance estimation* (CVE). CVE falls in the class of nonparametric methods. The estimators in this class minimise 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 span$\{\mathbf{B}\}$. Statistical approaches to identify $\mathbf{B}$ in (2) include ordinary least squares and nonparametric multiple index models. The OLS estimator, $\mathbf{\Sigma}_\mathbf{X}^{-1}\text{cov}(\mathbf{X}, Y)$, always falls in span$\{\mathbf{B}\}$ [see Theorem 8.3, Li (2018)]. Principal Hessian Directions (pHd, Li (1992)) was the first SDR estimator to target 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 (Cook and Li, 2004), still requires the linearity condition in order to recover vectors in span($\mathbf{B}$).

The most competitive nonparametric SDR method up to now has been the minimum average variance estimation method (MAVE, Xia et al. (2002)). MAVE assumes model (1), bounded fourth derivative covariate density, and existence of continuous bounded third derivatives for g. It is based on a local first order approximation of g in (1) and the minimisation of the expected conditional variance of the response given $\mathbf{B}^\top \mathbf{X}$.

The conditional variance estimator (CVE) also targets and recovers span$\{\mathbf{B}\}$ 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 $\mathbf{X}$ should exhibit significantly higher variation in Y as compared with the directions along which Y exhibits markedly less

variation. CVE is a fully data-driven estimator that performs better or on par with MAVE in simulations. Furthermore, in contrast to MAVE, CVE does not estimate the link function g and requires weaker assumptions on its smoothness.

The rest of the paper is organised as follows. In Section 2 we define the proposed conditional variance estimator (CVE) and provide its geometrical motivation. Section 3 proposes the relevant estimators. The estimation optimization algorithm is given in Section 4. Statistical properties of the estimators are obtained in Section 5. Simulation studies are carried out in Section 6 and the Hitters data set is analysed in Section 7. We conclude in Section 8.

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 $\text{supp}(f_{\mathbf{X}})$. In the sequel, we refer to the following assumptions as needed.

Assumption A.1. *Model (1) holds with $g : \mathbb{R}^k \to \mathbb{R}$ non constant in all arguments, $\mathbf{X}$ stochastically independent from ϵ, $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}\text{ar}(\epsilon) = \eta^2 < \infty$, and $\mathbf{\Sigma}_{\mathbf{X}}$ is positive definite.*

Assumption A.2. *The link function g is continuous and $f_{\mathbf{X}}$ is continuous.*

Assumption A.3. $\mathbb{E}(|Y|^4) < \infty$.

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

Assumption A.5. $|Y| < M_2 < \infty$ *almost surely.*

The set

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

denotes a Stiefel manifold that comprises of all $p \times q$ matrices with orthonormal columns. $S(p, q)$ is compact and $\dim(S(p, q)) = pq - q(q + 1)/2$ [see W.M.Boothby (2002) and Section 2.1 of Tagare (2011)]. For $q \leq p \in \mathbb{N}$ and any $\mathbf{V} \in S(p, q)$, we define

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mathbb{V}\text{ar}(Y | \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \tag{4}$$

where $\mathbf{s}_0 \in \mathbb{R}^p$ is a shifting point. Since $\mathbf{X}$ has a continuous distribution, the set $\{\omega \in \Omega : \mathbf{X}(\omega) \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\}$ has probability 0 if $q < p$. Let

$$f_{\mathbf{X}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}}(\mathbf{x}) = \begin{cases} \frac{f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}} & \text{if } \mathbf{x} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}, \mathbf{r}_1 = \mathbf{V}^\top(\mathbf{x} - \mathbf{s}_0) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Theorem 1 establishes that (5) is a proper density and that $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (4), and its generalised version,

$$L(\mathbf{V}) = \int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right), \tag{6}$$

are well-defined using the concept of regular conditional probability (Leao Jr and et al., 2004). Moreover, Theorem 1 provides its formula.

Theorem 1. *Let* $\mathbf{X}$ *be a* p*-dimensional continuous random vector with density* $f_{\mathbf{X}}(\mathbf{x})$*. Under assumption A.2, for* $\mathbf{s}_0 \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$ *and* $\mathbf{V} \in S(p,q)$ *defined in (3), (5) is a proper density. Under assumptions A.1, A.2 and A.4, (4) and (6) are well defined and continuous for* $\mathbf{V} \in S(p,q)$ *and* $\mathbf{s}_0 \in supp(f_{\mathbf{X}})$*. Moreover,*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 + \eta^2 \tag{7}$$

where

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

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

Theorem 2 provides the statistical motivation for the objective function (6) of the conditional variance estimator.

Theorem 2. *Under assumptions A.1, A.2 and A.4,*

(a) *For all* $\mathbf{s}_0 \in \mathbb{R}^p$ *and* $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q)$ *such that there exist* $u \in \{1, ..., q\}$ *with* $\mathbf{v}_u \in \text{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \mathbb{V}\text{ar}(\epsilon) = \eta^2$.
(b) *For all* $\mathbf{s}_0 \in \mathbb{R}^p$ *and* $\mathbf{V} \in \text{span}\{\mathbf{B}\}^\perp$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \eta^2$.
(c) *For all* $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q)$ *such that there exist* $u \in \{1, ..., q\}$ *with* $\mathbf{v}_u \in \text{span}\{\mathbf{B}\}$, $L(\mathbf{V}) > \eta^2$.
(d) *For all* $\mathbf{V} \in \text{span}\{\mathbf{B}\}^\perp$, $L(\mathbf{V}) = \eta^2$

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 $\mathbf{v}_u \in \text{span}\{\mathbf{B}\}$ for some $u \in \{1, ..., q\}$. To obtain (a), observe (4) yields

$$\begin{aligned} \tilde{L}(\mathbf{V}, \mathbf{s}_0) &= \mathbb{V}\text{ar}\big(g(\mathbf{B}^\top \mathbf{X}) | \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \mathbf{s}_0)\big) + \mathbb{V}\text{ar}(\epsilon) \\ &= \mathbb{V}\text{ar}\big(g(\mathbf{B}^\top \mathbf{s}_0 + \mathbf{B}^\top \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \mathbf{s}_0)) | \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \mathbf{s}_0)\big) + \eta^2 > \eta^2 \end{aligned} \tag{8}$$

since $\mathbf{B}^\top \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \mathbf{s}_0) \neq 0$ w.p. 1, and therefore the first term in (8) has positive variance. For $\mathbf{V}$ such that $\mathbf{V}$ and $\mathbf{B}$ are orthogonal, $\mathbf{B}^\top \mathbf{V}\mathbf{V}^\top(\mathbf{X} - \mathbf{s}_0) = 0$ and (b) follows. Since $\mathbf{s}_0$ is arbitrary yet constant, (c) and (d) follow. □

Theorem (2) also has a geometrical motivation. If $\mathbf{X}$ is not random, the deterministic function $Y = g(\mathbf{B}^\top \mathbf{X})$ 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 ϵ in all directions orthogonal to $\mathbf{B}$. In all other directions the variation comprises of the sum of the variation of ϵ and of $g(\mathbf{B}^\top \mathbf{X})$. In consequence, the objective function (6) captures the variation of Y as $\mathbf{X}$ varies in the column space of $\mathbf{V}$ and is minimised in the directions orthogonal to $\mathbf{B}$.

2.1 | Conditional Variance Estimator (CVE)

The objective function $L(\mathbf{V})$ is well defined and continuous by Theorem 1. Let

$$\mathbf{V}_q = \text{argmin}_{\mathbf{V} \in S(p,q)} L(\mathbf{V}). \tag{9}$$

$\mathbf{V}_q$ is well defined as the minimiser of a continuous function over the compact set $S(p,q)$. Corollary 3 follows directly from Theorem 2 and provides the means for identifying the linear projections of the predictors satisfying (1).

Corollary 3. *Under the assumptions of Theorems 1 and 2, the solution of the optimisation problem in (9) is well defined and*

(a) $\operatorname{span}\{\mathbf{V}_{p-k}\} = \operatorname{span}\{\mathbf{B}\}^{\perp}$
(b) $\operatorname{span}\{\mathbf{V}_{p-k}\}^{\perp} = \operatorname{span}\{\mathbf{B}\}$

where $k = \dim(\operatorname{span}\{\mathbf{B}\})$.

The minimiser $\mathbf{V}_{p-k}$ in Corollary 3 is not unique since for all $\mathbf{C} \in \mathbb{R}^{q\times p-k}$ such that $\mathbf{C}\mathbf{C}^{\top} = \mathbf{I}_{p-k}$, $L(\mathbf{V}\mathbf{C}) = L(\mathbf{V})$ since $L(\mathbf{V})$ depends on $\mathbf{V}$ only through $\operatorname{span}\{\mathbf{V}\}$. Nevertheless, since every minimiser spans the same subspace, $\operatorname{span}\{\mathbf{B}\}$ is uniquely identifiable.

Theorem 2 (c) and (d) lead to the proposed method for the identification of the sufficient reduction space, $\operatorname{span}\{\mathbf{B}\}$, in (1). Corollary 3 (b) serves as the estimation equation for CVE at the population level.

Definition 4. *The* ***Conditional Variance Estimator*** *is defined to be any basis of* $\operatorname{span}\{\mathbf{V}_q\}^{\perp}$, *where*

$$\mathbf{B}_{p-q} = \mathbf{V}_q^{\perp} \tag{10}$$

We can also target $\mathbf{B}$ directly by maximising the objective function $L(\mathbf{V})$. The downside of this approach is that $\mathbf{X}$ either needs to be standardised, or the conditioning argument needs to be changed to $\mathbf{X} = \mathbf{s}_0 + \mathbf{V}(\mathbf{V}^{\top}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\mathbf{V})^{-1}\mathbf{V}^{\top}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{X}-\mathbf{s}_0)$, or, equivalently, $\mathbf{X} = \mathbf{s}_0 + \mathbf{P}_{\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\operatorname{span}\{\mathbf{V}\})}(\mathbf{X}-\mathbf{s}_0)$, where $\mathbf{P}_{M(\operatorname{span}\{\mathbf{V}\})}$ is the orthogonal projection operator with respect to the inner product $\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{M}} = \mathbf{x}^{\top}\mathbf{M}\mathbf{y}$. In either case, the inversion of $\boldsymbol{\Sigma}_{\mathbf{X}}$ is required. Our choice of targeting the orthogonal complement avoids the inversion of $\boldsymbol{\Sigma}_{\mathbf{X}}$, and the method applies to regressions with $p > n$ or $p \approx n$.

3 | ESTIMATION OF $L(\mathbf{V})$

Assume $(Y_i, \mathbf{X}_i^{\top})_{i=1,\dots,n}^{\top}$ is an i.i.d. sample from model (1). We define

$$\begin{aligned} d_i(\mathbf{V}, \mathbf{s}_0) &= \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}}\mathbf{X}_i\|_2^2 = \|\mathbf{X}_i - \mathbf{s}_0\|_2^2 - \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{V}\mathbf{V}^{\top}(\mathbf{X}_i - \mathbf{s}_0)\rangle \\ &= \|(\mathbf{I}_p - \mathbf{V}\mathbf{V}^{\top})(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 = \|\mathbf{Q}_{\mathbf{V}}(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 \end{aligned} \tag{11}$$

where $\langle\cdot,\cdot\rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^{\top}$ and $\mathbf{Q}_{\mathbf{V}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{V}}$. Furthermore, let $h_n \in \mathbb{R}_+$ represent the width of a slice around the subspace $\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$ that satisfies $h_n \to 0$, $nh_n^{p-q} \to \infty$.

Let $K : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive, non increasing, monotone and bounded function (i.e. $|K(\cdot)| \le M_1$) with $\int_{\mathbb{R}^q} K(\|\mathbf{r}\|_2^2)d\mathbf{r} < \infty$ for $q \le p-1$, which we refer to as *kernel*. Examples of such functions include the rectangular, $K(z) = cI(z \le 1)$, the Gaussian, $K(z) = c\exp(-z^2/2)$, the exponential, $K(z) = c\exp(-z)$, and the Epanechnikov kernel, $K(z) = c\max\{(1-z^2), 0\}$, where c is a constant. A list of admissible kernel functions are given in (Parzen, 1961, Table 1). For $i = 1, \dots, n$, we let

$$w_i(\mathbf{V}, \mathbf{s}_0) = \frac{K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{d_j(\mathbf{V},\mathbf{s}_0)}{h_n}\right)} \tag{12}$$

The sample based estimate of $\tilde{L}(\mathbf{V}, s_0)$ is defined as

$$\tilde{L}_n(\mathbf{V}, s_0) = \sum_{i=1}^{n} w_i(\mathbf{V}, \mathbf{s}_0)(Y_i - \bar{y}_1(\mathbf{V}, \mathbf{s}_0))^2 = \bar{y}_2(\mathbf{V}, \mathbf{s}_0) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0)^2 \tag{13}$$

where $\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \sum_{i=1}^{n} w_i(\mathbf{V}, \mathbf{s}_0) Y_i^l$, $l = 1, 2$. The estimate of the objective function $L(\mathbf{V})$ in (6) is defined as

$$L_n(\mathbf{V}) = \frac{1}{n} \sum_{i=1}^{n} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \tag{14}$$

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

$L_n(\mathbf{V})$ in (14) depends on the weights $w_i(\mathbf{V}, \mathbf{s}_0)$ defined in (12). These are not only stochastically dependent but also random functions of the parameter $\mathbf{V}$, which is also the estimation target. This is novel in nonparametric estimation and poses challenges in obtaining theoretical properties of the estimator, as the standard probability tools do not apply.

To obtain insight as to the choice of $\tilde{L}_n(\mathbf{V}, s_0)$ in (13), we consider the rectangular kernel, $K(z) = 1_{\{z \leq 1\}}$. In this case, $\tilde{L}_n(\mathbf{V}, s_0)$ computes the empirical variance of the Y_i's corresponding to the $\mathbf{X}_i$'s that are no further than h_n away from the subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, $\|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0 + \text{span}\{\mathbf{V}\}} \mathbf{X}_i\|_2^2 \leq h_n$. If a smooth kernel is used, such as the Gaussian in our simulation studies, then $\tilde{L}_n(\mathbf{V}, s_0)$ is also smooth, which allows the computation of gradients required to solve the optimization problem. We compute the gradient of (13) and (14) for the Gaussian kernel in Theorem 5, which is proven in the Appendix.

Theorem 5. *The gradient of $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ in (13) 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 (14) is

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

3.1 | Weighted estimation of $L(\mathbf{V})$

We call the set $\mathcal{S}_{\mathbf{s}_0, \mathbf{V}} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0 + \text{span}\{\mathbf{V}\}} \mathbf{x}\|_2^2 \leq h_n\}$ a *slice* that depends on both the shifting point $\mathbf{s}_0$ and $\mathbf{V}$. In the estimation of $L(\mathbf{V})$ two different weighting schemes are used:

(a) Within $\mathcal{S}_{\mathbf{s}_0, \mathbf{V}}$ (within a slice). The formula is given by (12) and it is used to calculate (13)

(b) Between $\mathcal{S}_{\mathbf{s}_0, \mathbf{V}}$ (between different slices). Here equal weights $1/n$ are used to calculate (14)

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 realised by altering (14) to:

$$L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^{n} \tilde{w}(\mathbf{V}, \mathbf{X}_i)\tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \tag{15}$$

$$\tilde{w}(\mathbf{V}, \mathbf{X}_i) = \frac{\sum_{j=1}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n) - 1}{\sum_{l,u=1}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n) - n} = \frac{\sum_{j=1, j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)}{\sum_{l,u=1, l\neq u}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n)} \tag{16}$$

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, i.e. the higher the number of observations that we use for estimating $L(\mathbf{V}, \mathbf{X}_i)$ the higher the accuracy of this estimation. The denominator in (16) guarantees the weights $\tilde{w}(\mathbf{V}, \mathbf{X}_i)$ sum up to one.

Theorem 6. *The gradient of $L_n^{(w)}(\mathbf{V})$ in (15) is given by*

$$\nabla_{\mathbf{V}} L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^{n} \left(\nabla_{\mathbf{V}} \tilde{w}(\mathbf{V}, \mathbf{X}_i)\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) + \tilde{w}(\mathbf{V}, \mathbf{X}_i)\nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)\right), \tag{17}$$

where $\nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$ is given in (5) and if $K(\cdot)$ is the Gaussian kernel,

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

with $K_{j,i} = K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$ and $d_{j,i} = d_j(\mathbf{V}, \mathbf{X}_i)$.

If (15) and the gradient in (17) is used in the optimisation algorithm in Section 4, we refer to the estimator as *weighted CVE*. If (15) 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 (17) is dropped, we refer to it as *partially weighted CVE*. The weighted version of CVE is expected to increase the accuracy of the estimator for unevenly spaced data.

3.2 | Bandwidth selection

The performance of CVE depends crucially on the choice of the bandwidth sequence h_n that controls the bias-variance trade-off: 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 the bandwidth satisfies the following conditions:(a) $\lim_{n\to\infty} h_n = 0$ and (b) $\lim_{n\to\infty} n h_n^{p-q} = \infty$.

Theorem 7. *Let $\mathbf{M}$ be a $p \times p$ positive definite matrix. Then,*

$$\frac{tr(\mathbf{M})}{p} = \text{argmin}_{s>0} \|\mathbf{M} - s\mathbf{I}_p\|_2 \tag{18}$$

Proof. Let $\mathbf{U}$ be the $p \times p$ matrix whose columns are the eigenvectors of $\mathbf{M}$ corresponding to its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p > 0$. Then, $\mathbf{M} = \mathbf{U}\text{diag}(\lambda_1, ..., \lambda_p)\mathbf{U}^\top$, which implies $\|\mathbf{M} - s\mathbf{I}_p\|_2^2 = \|\text{diag}(\lambda_1, ..., \lambda_p) - s\mathbf{I}_p\|_2^2 = \sum_{l=1}^{p} (\lambda_l - s)^2$. Taking the derivative with respect to s, setting it to 0 and solving for s obtains (18), since $\sum_{l=1}^{p} \lambda_l = \text{tr}(\mathbf{M})$. □

In order to avoid bandwidth dependence on $\mathbf{V}$, we assume the predictors are multivariate normal, so that their

joint density is approximated by $N(\boldsymbol{\mu}_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$ by Theorem 7, for $\sigma^2 = \text{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})/p$. Under $\mathbf{X} \sim N_p(\boldsymbol{\mu}_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$, $\tilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbf{X}_j \sim N_p(0, 2\sigma^2\mathbf{I}_p)$ for $i \neq j$, where we suppress the dependence on j for notational convenience. Since all data are used as shifting points, $d_i(\mathbf{V}, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 - (\mathbf{X}_i - \mathbf{X}_j)^\top\mathbf{V}\mathbf{V}^\top(\mathbf{X}_i - \mathbf{X}_j) = \|\tilde{\mathbf{X}}_i\|_2^2 - \tilde{\mathbf{X}}_i^\top\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{X}}_i$. Let

$$\begin{aligned}\text{nObs} &= \mathbb{E}(\#\{i \in \{1, \ldots, n\} : \tilde{\mathbf{X}}_i \in \text{span}_h\{\mathbf{V}\}\}) \\ &= 1 + (n-1)\mathbf{P}(d_1(\mathbf{V}, \mathbf{X}_2) \leq h) = 1 + (n-1)\mathbf{P}(\|\tilde{\mathbf{X}}\|_2^2 - \tilde{\mathbf{X}}^\top\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{X}} \leq h)\end{aligned} \tag{19}$$

where $\text{span}_h\{\mathbf{V}\} = \{\mathbf{x} \in \mathbf{R}^p : \|\mathbf{x} - \mathbf{P}_{\text{span}\{\mathbf{V}\}}\mathbf{x}\|_2^2 \leq h\}$ and $\tilde{\mathbf{X}} = \mathbf{X} - \tilde{\tilde{\mathbf{X}}}$, with $\tilde{\tilde{\mathbf{X}}}$ 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 (19).

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

$$\mathbf{P}(\|\tilde{\mathbf{X}}\|_2^2 - \tilde{\mathbf{X}}^\top\mathbf{V}\mathbf{V}^\top\tilde{\mathbf{X}} \leq h) = \mathbf{P}(\|\mathbf{R}_2\|_2^2 \leq h) = \chi_{p-q}\left(\frac{h}{2\sigma^2}\right), \tag{20}$$

where χ_{p-q} is the cdf of a chi-squared distribution with $p - q$ degrees of freedom. Plugging (20) in (19) obtains

$$\text{nObs} = 1 + (n-1)\chi_{p-q}\left(\frac{h}{2\sigma^2}\right). \tag{21}$$

Solving (21) for h and Theorem 7 yield

$$h_n(\text{nObs}) = \chi_{p-q}^{-1}\left(\frac{\text{nObs} - 1}{n-1}\right)\frac{2\text{tr}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})}{p}, \tag{22}$$

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

In order to ascertain h_n satisfies conditions (a) and (b) in the beginning of this section, 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.

$$h_n = 1.2^2\frac{2\text{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}})}{p}\left(n^{-1/(4+p-q)}\right)^2 \tag{23}$$

where $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}$ is the maximum likelihood estimate of $\boldsymbol{\Sigma}_{\mathbf{X}}$. The term $2\text{tr}(\hat{\boldsymbol{\Sigma}}_{\mathbf{X}})/p$ can be interpreted as the variance of $\mathbf{X}_i - \mathbf{X}_j$ and $p - q$ stands for the true dimension k. We use 1.2 as c based on empirical evidence from the simulations. Since both (22) and (23) yield satisfactory results, we opted against cross validation because of the computational burden involved.

4 | OPTIMIZATION ALGORITHM

A Stiefel manifold optimization algorithm is used to obtain the solution of the sample version of the optimization problem (9). To calculate $\widehat{\mathbf{V}}_q$ in (6) a curvilinear search is used (Zaiwen Wen, 2013; Tagare, 2011), an approach similar to gradient descend. First an arbitrary starting value $\mathbf{V}^{(0)}$ is selected by drawing a $p \times q$ matrix from the invariant measure on $S(p,q)$; i.e., the uniform distribution on $S(p,q)$. The Q-component of the QR decomposition of a $p \times q$ matrix with independent standard normal entries follows the invariant measure (Chikuse, 1994). The step-size $\tau > 0$ and tolerance tol > 0 are fixed at the outset.

Result: $\mathbf{V}^{(\text{end})}$
Initialise: $\mathbf{V}^{(0)}$, $\tau = 1$, tol $= 10^{-3}$, $\gamma = 0.5$ error = tol + 1, maxit = 50, count = 0;
while *error* $>$ *tol and count* $\leq$ *maxit* **do**
- $\mathbf{G} = \nabla_{\mathbf{V}} L_n(\mathbf{V}^{(j)}) \in \mathbb{R}^{p\times q}$, $\mathbf{W} = \mathbf{G}\mathbf{V}^\top - \mathbf{V}\mathbf{G}^\top \in \mathbb{R}^{p\times p}$
- $\mathbf{V}^{(j+1)} = (\mathbf{I}_p + \tau\mathbf{W})^{-1}(\mathbf{I}_p - \tau\mathbf{W})\mathbf{V}^{(j)}$
- error $= \|\mathbf{V}^{(j)}\mathbf{V}^{(j)\top} - \mathbf{V}^{(j+1)}\mathbf{V}^{(j+1)^\top}\|_2/\sqrt{2q}$

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

Zaiwen Wen (2013) showed that the sequence generated by the algorithm converges to a stationary point if Armijo-Wolfe conditions Arm (2006) are used for determining the stepsize τ. We use simpler conditions to determine the step size since they are computationally less expensive and exhibit same behavior as the Armijo-Wolfe conditions in the simulations.

The algorithm is repeated for m arbitrary $\mathbf{V}^{(0)}$ starting values drawn from the invariant measure on $S(p,q)$. Among those, the value at which L_n in (14) is minimal is selected as $\widehat{\mathbf{V}}_q$.

5 | THEORY

In this section we show that the sample based objective function is weakly consistent for its true value. All proofs are given in the Appendix.

The summands of $\tilde{L}_n$ in (13) can be expressed as

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)}, \tag{24}$$

where

$$t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = \frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^{n} K(d_i(\mathbf{V}, \mathbf{s}_0)/h_n) Y_i^l \tag{25}$$

for $l = 0, 1, 2$.

Theorem 8. *Under assumptions A.1, A.3, and* $nh_n^{p-q} \to \infty$,

$$\mathbb{V}\mathrm{ar}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) \to 0$$

for $t_n^{(l)}$ *given in* (25), $l = 0, 1, 2$.

Theorem 9. *Under assumptions A.1, A.2, A.4, and* $h_n \to 0$,

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) g(\mathbf{B}^\top \mathbf{X}_i)^l\right) \to t^{(l)}(\mathbf{V}, \mathbf{s}_0) \int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}\|_2^2) d\mathbf{r}, \tag{26}$$

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) \epsilon_i\right) = 0, \tag{27}$$

and

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) \epsilon_i^2\right) \to \eta^2 t^{(0)}(\mathbf{V}, \mathbf{s}_0) \int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}\|_2^2) d\mathbf{r} \tag{28}$$

where $t^{(l)}$ *is defined in Theorem 1 for* $l = 0, 1, 2$.

Theorem 10. *Under assumptions A.1, A.2, A.3, A.4,* $h_n \to 0$, $nh_n^{p-q} \to \infty$ *and* $\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}\|_2^2) d\mathbf{r} = 1$,

(a) $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) \xrightarrow{L^2(\Omega)} t^{(l)}(\mathbf{V}, \mathbf{s}_0)$, *for* $l = 0, 1$
(b) $t_n^{(2)}(\mathbf{V}, \mathbf{s}_0) \xrightarrow{L^2(\Omega)} t^{(2)}(\mathbf{V}, \mathbf{s}_0) + \eta^2 t^{(0)}(\mathbf{V}, \mathbf{s}_0)$

for $t_n^{(l)}$ *given in* (25) *and* $t^{(l)}$ *defined in Theorem 1, for* $l = 0, 1, 2$.

Theorem 10 follows directly from Theorems 8, 9 and the bias variance decomposition,

$$\mathbb{E}(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t^{(l)}(\mathbf{V}, \mathbf{s}_0))^2 = \left(\mathbb{E}(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)) - t^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)^2 + \mathbb{V}\mathrm{ar}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right).$$

Theorem 11. *Under A.1, A.2, A.3, A.4,* $h_n \to 0$ *and* $nh_n^{p-q} \to \infty$,

(a) $\bar{y}_1(\mathbf{V}, \mathbf{s}_0) \xrightarrow{\mathbf{P}} \mu_1(\mathbf{V}, \mathbf{s}_0)$
(b) $\bar{y}_2(\mathbf{V}, \mathbf{s}_0) \xrightarrow{\mathbf{P}} \mu_2(\mathbf{V}, \mathbf{s}_0) + \eta^2$
(c) $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) \xrightarrow{\mathbf{P}} \tilde{L}(\mathbf{V}, \mathbf{s}_0)$

where $\bar{y}_l(\cdot,\cdot)$ is given in (24) and $\mu_l(\cdot,\cdot)$ in Theorem 1 for $l = 1, 2$.

Theorems 8-11 lead to Theorem 12 that establishes the consistency of the sample CVE objective function.

Theorem 12. *Under A.1, A.2, A.3, A.4, A.5, $h_n \to 0$ and $nh_n^{p-q} \to \infty$, then $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) \xrightarrow{L^2(\Omega)} \tilde{L}(\mathbf{V}, \mathbf{s}_0)$, and*

$$L_n(\mathbf{V}) \longrightarrow L(\mathbf{V}) \quad \textit{in probability}$$

as $n \to \infty$ for all $\mathbf{V} \in S(p, q)$.

5.1 | A small study of the behaviour of $L_n(\mathbf{V})$

We explore how accurately the sample version (14) of the objective function estimates the target subspace in an example. We consider a bivariate normal predictor vector, $\mathbf{X} = (X_1, X_2)^\top \sim N(0, \mathbf{\Sigma_X})$. We generate the response from $Y = g(\mathbf{B}^\top \mathbf{X}) + \epsilon = X_1 + \epsilon$, with $\epsilon \sim N(0, \eta^2)$ independent of $\mathbf{X}$. Therefore, $k = 1$, $\mathbf{B} = (1, 0)^\top$, $g(z) = z \in \mathbb{R}$ in (1).

Applying Theorem 1 obtains

$$\mu_l(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^2} g(\mathbf{B}^\top \mathbf{x})^l f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^2} (\mathbf{B}^\top \mathbf{x})^l f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) d\mathbf{x} \tag{29}$$

In the Appendix we show that, under this setting, (5) is given by

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) = \begin{cases} \frac{1}{\sigma}\psi(\frac{r_1-\alpha}{\sigma}) & \text{if } \mathbf{x} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}, r_1 = \mathbf{V}^\top(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R} \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

where $\psi(z)$ is the density of a standard normal variable. Inserting (30) in (29) yields

$$\int_{\mathbb{R}} (\mathbf{B}^\top \mathbf{s}_0 + \mathbf{B}^\top \mathbf{V} r_1)^l \frac{1}{\sigma}\psi(\frac{r_1 - \alpha}{\sigma}) dr_1 = \begin{cases} \mathbf{B}^\top \mathbf{s}_0 + \mathbf{B}^\top \mathbf{V}\alpha & l = 1 \\ (\mathbf{B}^\top \mathbf{s}_0)^2 + 2(\mathbf{B}^\top \mathbf{s}_0)(\mathbf{B}^\top \mathbf{V})\alpha + (\mathbf{B}^\top \mathbf{V})^2(\sigma^2 + \alpha^2) & l = 2 \end{cases}$$

for $\mathbf{V} \in \mathbb{R}^{2\times 1}$, $\sigma^2 = (\mathbf{V}^\top \mathbf{\Sigma}_\mathbf{X}^{-1} \mathbf{V})^{-1}$ and α is computed in the Appendix. Applying Theorem 1, using the definitions (4) and (6), yields $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 + \eta^2 = (\mathbf{B}^\top \mathbf{V})^2\sigma^2 + \eta^2$, so that

$$L(\mathbf{V}) = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right) = (\mathbf{B}^\top \mathbf{V})^2\sigma^2 + \eta^2 = \frac{(\mathbf{B}^\top \mathbf{V})^2}{\mathbf{V}^\top \mathbf{\Sigma}_\mathbf{X}^{-1} \mathbf{V}} + \eta^2 \tag{31}$$

From (31) we can easily see that $L(\mathbf{V})$ attains its minimum when $\mathbf{V} \perp \mathbf{B}$. Also, if $\mathbf{\Sigma_X} = \mathbf{I}_2$, the maximum of $L(\mathbf{V})$ is attained at $\mathbf{V} = \mathbf{B}$. To visualise the behaviour of $\tilde{L}_n(\mathbf{V})$ as the sample size increases, we parametrise $\mathbf{V}$ by $\mathbf{V}(\theta) = (\cos(\theta), \sin(\theta))^\top$, $\theta \in [0, \pi]$. Since $\mathbf{B} = (1, 0)^\top$, the minimum of $\tilde{L}(\mathbf{V})$ is at $\mathbf{V}(\pi/2) = (0, 1)^\top$, 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 size n in Fig 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 the smallest sample of 10 observations.

Assumptions A.4 and A.5 are violated in this example, which suggests that the proposed estimator of CVE applies under weaker assumptions.

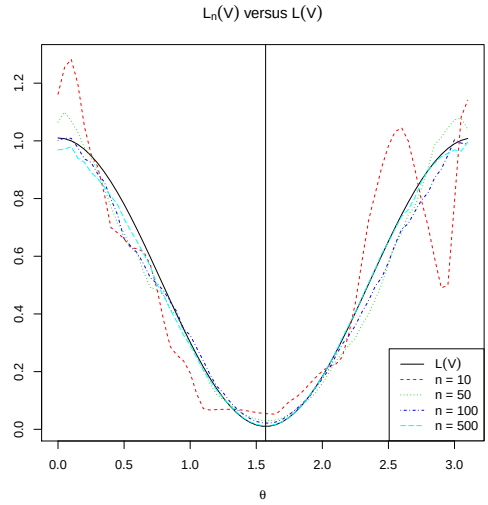

FIGURE 1 Solid black line is $L(\mathbf{V}(\theta)) = \cos(\theta)^2 + 0.1^2$, colored is $L_n(\mathbf{V}(\theta))$, $\theta \in [0, \pi]$, $n = 10, 50, 100, 500$. The vertical black line is at $\theta = \pi/2$

6 | SIMULATION STUDIES

We compare the estimation accuracy of CVE with the forward model based SDR methods, mean OPG (`meanOPG`), mean MAVE (`meanMAVE`) (Weiqiang and Yingcun, 2019), rOPG (`rOPG`), rmave (`rmave`) (Xia et al., 2002; Li, 2018), and `pHd` (Li, 1992; Cook and Li, 2002), and the inverse regression based methods, SIR Li (1991) and SAVE Cook and Weisberg (1991). The dimension k is assumed to be known throughout.

We report results for CVE using the "plug-in" bandwidth in (23) and three different CVE versions. CVE is obtained by using $m = 10$ arbitrary starting values in the optimization algorithm and optimizing (14) as described in Section 4. rCVE, or *refined weighted CVE*, is obtained by using one starting value, $V^{(0)}$ equal to the optimiser of CVE, and using (15) in the optimization algorithm in Section 4, with the partially weighted gradient as described in Section 3.1. wCVE, or *weighted CVE*,is obtained by optimizing (15) with partially weighted gradient as described in sections 3.1 and 4. Methods rOPG and rMAVE refer to the original refined OPG and refined MAVE algorithms published in Xia et al. (2002). They are implemented using the `R` code in Li (2018) with nit = 25 (number of iterations is 25 since empirically the algorithm seems to have converged). The `dr` package is used for the SIR, SAVE and pHd calculations, and the `MAVE` package for mean OPG and mean MAVE. The source code for CVE can be found at `https://git.art-ist.cc/daniel/CVE`. For M7 the results of `rOPG` and `rmave` are not reported because the code frequently produces an error message that a matrix is not invertible.

Table 1 lists the seven models (M1-M7) we consider. Throughout, we set $p = 20$, $\mathbf{b}_1 = (1, 1, 1, 1, 1, 1, 0, \ldots, 0)^\top/\sqrt{6}$, $\mathbf{b}_2 = (1, -1, 1, -1, 1, -1, 0, \ldots, 0)^\top/\sqrt{6} \in \mathbb{R}^p$ 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 like in M6 and $\mathbf{b}_4 = \mathbf{e}_3$ where $\mathbf{e}_j$ denotes the j unit vector in $\mathbb{R}^p$. The error term ϵ is independent of $\mathbf{X}$ for all models. In M2, M3, M4, M5 and M6, $\epsilon \sim N(0, 1)$. For M1 and M7, ϵ is distributed as generalised normal distribution Nadarajah (2005) with location 0, shape-parameter 0.5 for M1, and shape-parameter 1 for M7 (Laplace distribution), and for both the scale-parameter is chosen such that $\mathbb{V}\mathrm{ar}(\epsilon) = 0.25$.

TABLE 1 Models

Name	Model	**X** distribution	k	n
M1	$Y = \cos(\mathbf{b}_1^\top \mathbf{X}) + \epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	1	100
M2	$Y = \cos(\mathbf{b}_1^\top \mathbf{X}) + 0.5\epsilon$	$\mathbf{X} \sim Z1_p\lambda + N_p(0, \mathbf{I}_p)$	1	100
M3	$Y = 2\log(\lvert\mathbf{b}_1^\top \mathbf{X}\rvert + 2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{I}_p)$	1	100
M4	$Y = (\mathbf{b}_1^\top \mathbf{X})/(0.5 + (1.5 + \mathbf{b}_2^\top \mathbf{X})^2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	2	200
M5	$Y = \cos(\pi \mathbf{b}_1^\top \mathbf{X})(\mathbf{b}_2^\top \mathbf{X} + 1)^2 + 0.5\epsilon$	$\mathbf{X} \sim U([0,1]^p)$	2	200
M6	$Y = (\mathbf{b}_1^\top \mathbf{X})^2 + (\mathbf{b}_2^\top \mathbf{X})^2 + (\mathbf{b}_3^\top \mathbf{X})^2 + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{I}_p)$	3	200
M7	$Y = (\mathbf{b}_1^\top \mathbf{X})(\mathbf{b}_2^\top \mathbf{X})^2 + (\mathbf{b}_3^\top \mathbf{X})(\mathbf{b}_4^\top \mathbf{X}) + \epsilon$	$\mathbf{X} \sim t_3(\mathbf{I}_p)$	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, $\mathbf{X} \sim U([0,1]^p)$; i.e., 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 Xia et al. (2002), 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 $1_q = (1, 1, ..., 1)^\top \in \mathbb{R}^q$, 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_{mix} is the relative mode height and λ 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 SDR methods. The accuracy of the estimates is assessed using $err = \|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|_2/\sqrt{2k} \in [0, 1]$, where $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}\mathbf{B}^\top)^{-1}\mathbf{B}^\top$ is the orthogonal projection matrix on $\text{span}\{\mathbf{B}\}$. The factor $\sqrt{2k}$ normalises the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement.

The box-plots of the $r = 100$ estimation errors for each method are displayed in Figures 2 - 5. In Table 2 the means and standard deviations of err for M1-M7 (for M2 $p_{mix} = 0.3$ and $\lambda = 1$) are reported For models M1, M5, and M7, CVE is approximately on par with MAVE, its main competitor, as can be seen in Figs 2 - 5. SIR and SAVE are not competitive throughout our experiments. SIR, in particular, is expected to fail in models M1-M3, and M6 since $\mathbb{E}(Y|\mathbf{X})$ is even.

CVE shows its advantage in M3 [see Figure 2] and in M2 [see Figure 5]. `meanOPG` and `meanMAVE` are slightly more accurate than CVE in M4 [see Figure 3] and M6 [see Figure 4]. In M1 [see Figure 2], M5 [see Figure 3], and M7 [see Figure 4] CVE, `meanOPG` and `meanMAVE` are roughly on par. In M1, M3, M5, and M6 we observe a discrepancy between `meanOPG`, `meanMAVE` and `rOPG`, `rmave`; that is, the raw implementation of the refined OPG and MAVE algorithms perform worse than the implementation in the `R` package.

In Fig. 5, 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 but the reference methods are restricted to meanOPG and meanMAVE, since the others are not competitive. CVE 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_{mix}. The performance of both `meanOPG` and `meanMAVE` worsens as one moves from left to right row-wise. The mixing probability, p_{mix}, has no noticeable effect on the performance of any method; i.e., the plots are very similar column-wise. In sum, MAVE's performance deteriorates as the bimodality of the predictor distribution becomes more distinct. In contrast, CVE is unaffected.and appears to have an advantage over MAVE 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. CVE is the only method that estimates the mean subspace reliably in model M2 ($err \approx 0.4$ to 0.5), whereas MAVE misses it completely ($err \approx 1$).

These results indicate that CVE is often approximately on par, and can perform much better than MAVE 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
M1 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
M2 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
M3 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
M4 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
M5 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
M6 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
M7 sd	0.0588	0.0862	0.0879	0.1122	NA	0.1271	NA	0.0354	0.0341	0.0342

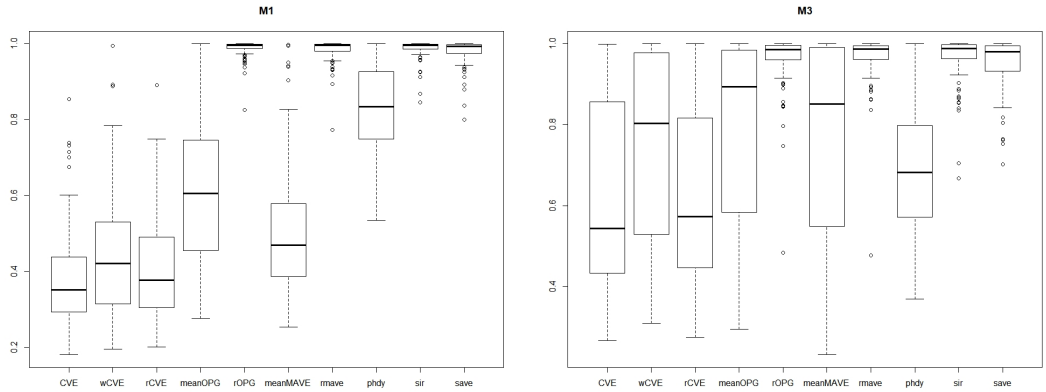

FIGURE 2 Left panel: M1, $p = 20, n = 100$; Right panel: M3, $p = 20, n = 100$

Furthermore we estimate the dimension k via cross-validation, following the approach in Xia et al. (2002) , with

$$\hat{k} = \text{argmin}_{l=1,\ldots,p} CV(l), \tag{32}$$

where $CV(l) = \sum_i (Y_i - \hat{g}^{-i}(\widehat{\mathbf{B}}_l^\top \mathbf{X}_i))^2/n$, $\hat{g}^{-i}(\cdot)$ is estimated from the dataset $(Y_j, \widehat{\mathbf{B}}_l^\top \mathbf{X}_j)_{j=1,\ldots,n;j\neq i}$ using multivariate adaptive regression splines (Friedman, 1991) implemented in the `R`-package `mda`, and $\widehat{\mathbf{B}}_l = \widehat{\mathbf{V}}_{p-l}^{\perp}$ is any basis of the

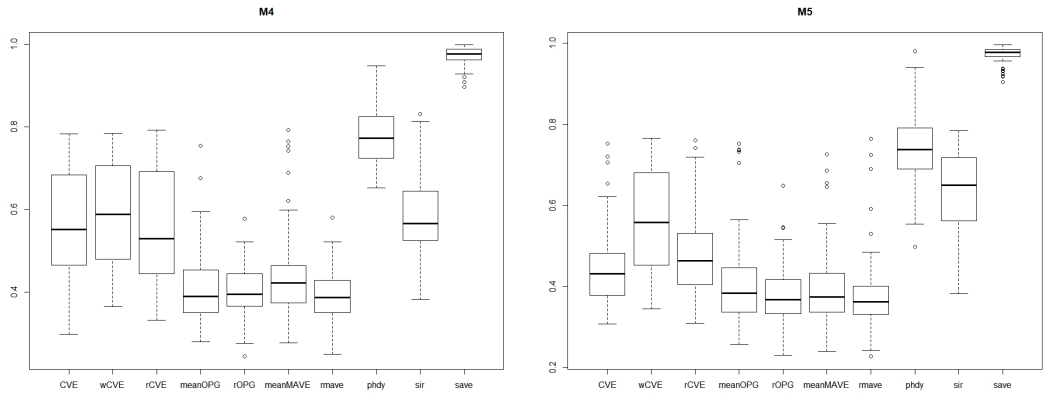

FIGURE 3 Left panel: M4, $p = 20, n = 200$; Right panel: M5 $p = 20, n = 200$

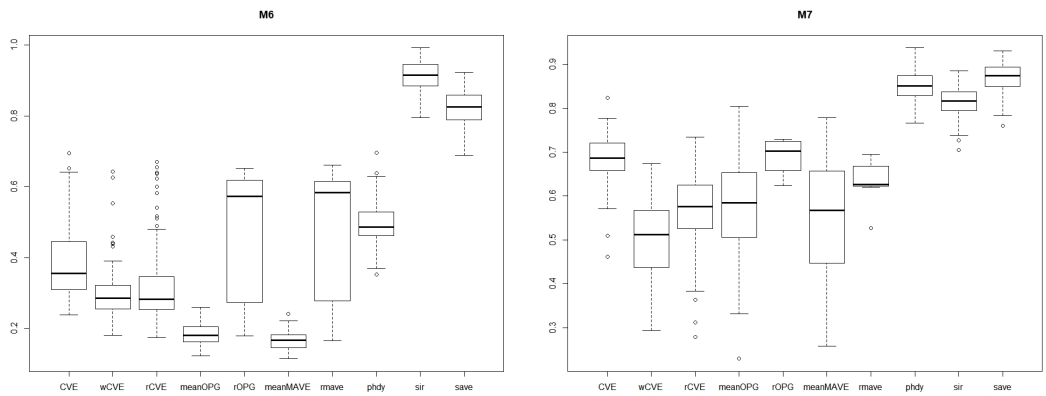

FIGURE 4 Left panel: M6, $p = 20, n = 200$; Right panel: M7 $p = 20, n = 400$

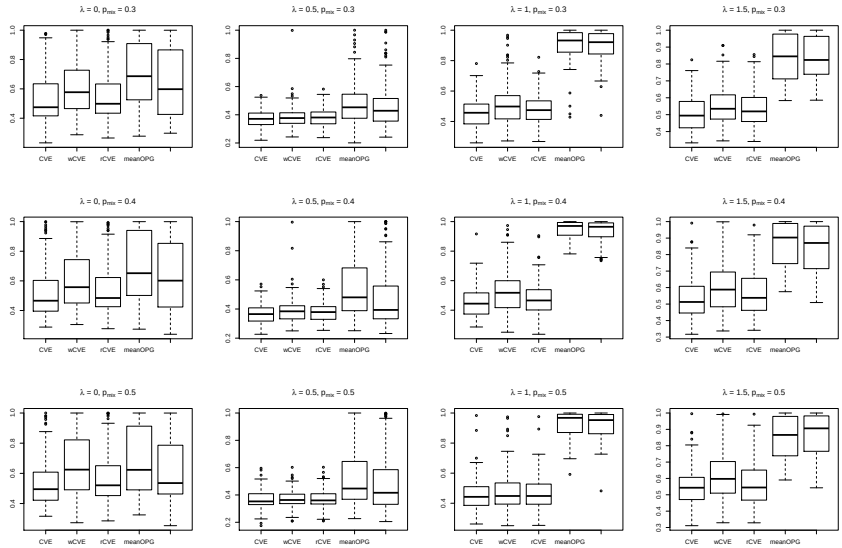

FIGURE 5 M2, $p = 20, n = 100$

orthogonal complement of $\widehat{\mathbf{V}}_{p-l}$, with

$$\widehat{\mathbf{V}}_{p-l} = \text{argmin}_{\mathbf{V} \in S(p,p-l)} L_n(\mathbf{V}).$$

For a given l, we calculate $\widehat{\mathbf{B}}_l$ from the whole data set and predict Y_i by $\hat{Y}_{i,l} = \hat{g}^{-i}(\widehat{\mathbf{B}}_l^\top \mathbf{X}_i)$. For $l = p$, $\widehat{\mathbf{B}}_p = \mathbf{I}_p$. The results for the seven models are reported in Table 3.

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

7 | HITTERS DATA SET

The Hitters data were analysed by Xia et al. (2002). The response is $Y = \log(\text{salary})$ and the covariate vector is the 16-dimensional $\mathbf{X} = (x_1, ..., x_{16})^\top$. Its components are times at bat x_1, hits x_2, home runs x_3, runs x_4, runs batted in x_5 and walks x_6 in 1986, years in major leagues x_7, times at bat x_8, hits x_9, home runs x_{10}, runs x_{11}, runs batted in x_{12} and walks x_{13} during their entire career up to 1986, put-outs x_{14}, assistances x_{15} and errors x_{16}. Following Xia et al. (2002), we standardise $\mathbf{X}$ by subtracting the mean and rescaling column-wise so that each predictor has unit variance. The same is done for Y. Furthermore, the 7 outliers are removed as in Xia et al. (2002).

TABLE 4 Mean cross-validation error

l	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

Table 4 reports the average cross validation mean squared error over $l = 1, \ldots, 5$. Both CVE and mean MAVE estimate the dimension to be 2.

Following Xia et al. (2002), we plot the response against the estimated directions in Fig. 6. CVE and MAVE pick

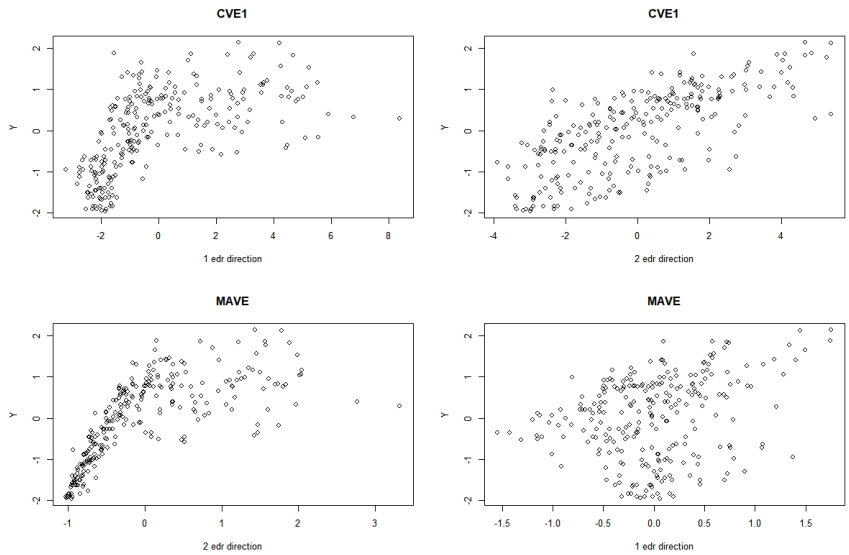

FIGURE 6 Y against $b_1^\top \mathbf{X}$ and $b_2^\top \mathbf{X}$

up the same pattern: the response appears to be linear in one direction and quadratic in the second.

Based on the scatterplots in Figure 6, we fit the same models for both the CVE and MAVE predictors. For CVE, the fitted regression is

$$\hat{Y} = 0.39578 + 0.33724(b_1^\top \mathbf{X}) - 0.08066(b_1^\top \mathbf{X})^2 + 0.29126(b_2^\top \mathbf{X}) \tag{33}$$

with $R^2 = 0.7975$, and for MAVE

$$\hat{Y} = 0.39051 + 1.32529(b_1^\top \mathbf{X}) - 0.55328(b_1^\top \mathbf{X})^2 + 0.49546(b_2^\top \mathbf{X}) \tag{34}$$

with $R^2 = 0.7859$. Both models (33) and (34) have about the same fit as measured by R^2.

8 | DISCUSSION

In this paper the novel conditional variance estimator (CVE) for the mean subspace is introduced. We present its geometrical and theoretical foundation and propose an estimation algorithm with assured convergence. CVE requires weak assumptions on the covariates, such as continuous density with compact support. The latter is sufficient but not necessary to show the sample objective function is consistent.

MAVE estimates the sufficient dimension reduction targeting both the reduction and the link function g in (1). CVE only targets the reduction and does not require estimation of the link function, which may explain why CVE has an advantage over MAVE, its direct competitor, in some regression settings. For example, in unreported simulations for model M2, CVE exhibits similar and better performance across different link functions (cos, exp, etc) for fixed λ, whereas the performance of MAVE is very uneven. Based on our simulations, it appears that when the link function is even and the predictor distribution is bimodal, CVE is more accurate than MAVE. 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 CVE arises from the novelty of its definition that involves random weights that depend on the parameter to be estimated. This precludes the usage of most standard probabilistic arguments for establishing consistency of the subspace estimates. A complete study of the asymptotic properties of CVE, optimal bandwidth selection and its extension to central space estimation are under investigation.

References

(2006) *Line Search Methods*, 30–65. New York, NY: Springer New York. URL: https://doi.org/10.1007/978-0-387-40065-5_3.

Adragni, K. and Cook, R. (2009) Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**, 4385–4405.

Bura, E., Duarte, S. and Forzani, L. (2016) Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, **111**, 1313–1329.

Bura, E. and Forzani, L. (2015) Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, **110**, 420–434.

Chikuse, Y. (1994) *Invariant measures on Stiefel manifolds with applications to multivariate analysis*, vol. Volume 24 of *Lecture Notes-Monograph Series*, 177–193. Hayward, CA: Institute of Mathematical Statistics. URL: https://doi.org/10.1214/lnms/1215463795.

Cook, D. R. (1998) *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. and Li, B. (2002) Dimension reduction for conditional mean in regression. *Ann. Statist.*, **30**, 455–474. URL: https://doi.org/10.1214/aos/1021379861.

Cook, R. D. (2007) Fisher lecture: Dimension reduction in regression. *Statist. Sci.*, **22**, 1–26.

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

— (2009) Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.

Cook, R. D. and Li, B. (2004) Determining the dimension of iterative hessian transformation. *Ann. Statist.*, **32**, 2501–2531. URL: https://doi.org/10.1214/009053604000000661.

Cook, R. D. and Weisberg, S. (1991) Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, **86**, 328–332. URL: http://www.jstor.org/stable/2290564.

Friedman, J. H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–67. URL: http://www.jstor.org/stable/2241837.

Heuser, H. (1995) *Analysis 2, 9 Auflage*. Teubner.

Leao Jr, D. and et al. (2004) Regular conditional probability, disintegration of probability and Radon spaces. *Proyecciones*, **23**, 15–29. URL: https://scielo.conicyt.cl/pdf/proy/v23n1/art02.pdf.

Li, B. (2018) *Sufficient dimension reduction: methods and applications with R*. CRC Press, Taylor & Francis Group.

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

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

Ma, Y. and Zhu, L. (2013) A review on dimension reduction. *International Statistical Review*, **81**, 134–150.

Nadarajah, S. (2005) A generalized normal distribution. *Journal of Applied Statistics*, **32**, 685–694. URL: https://doi.org/10.1080/02664760500079464.

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

Tagare, H. D. (2011) Notes on optimization on stiefel manifolds. URL: http://noodle.med.yale.edu/~hdtag/notes/steifel_notes.pdf.

Weiqiang, H. and Yingcun, X. (2019) *MAVE: Methods for Dimension Reduction*. URL: https://CRAN.R-project.org/package=MAVE. R package version 1.3.10.

W.M.Boothby (2002) *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002) An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 363–410. URL: http://dx.doi.org/10.1111/1467-9868.03411.

Zaiwen Wen, W. Y. (2013) A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, **142**, 397–434.

9 | APPENDIX

Proof of Theorem 1: The conditional probability of $\mathbf{X}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ is defined to be

$$\mathbf{P}(\mathbf{X} \leq \mathbf{x}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \lim_{h \downarrow 0} \frac{\mathbf{P}(\{\mathbf{X} \leq \mathbf{x}\} \cap \{\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}\})}{\mathbf{P}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\})} \tag{35}$$

where $\text{span}_h\{\mathbf{V}\} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \text{Pr}_{\text{span}\{\mathbf{V}\}}\mathbf{x}\|_2^2 \leq h\}$. Let $\mathbf{U}$ be an orthonormal basis of the orthogonal complement of $\text{span}\{\mathbf{V}\}$; that is, $\mathbf{U}^\top\mathbf{V} = 0, \mathbf{U}^\top\mathbf{U} = \mathbf{I}_{p-q}$. Let $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$ where $\mathbf{r}_1 = \mathbf{V}^\top(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^q, \mathbf{r}_2 = \mathbf{U}^\top(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^{p-q}$.

Then,

$$\begin{aligned}\mathbf{P}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}) &= \int_{\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \int_{\text{span}_h\{\mathbf{V}\}} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{x})d\mathbf{x} \\ &= \int_{\mathbb{R}^q}\int_{\|\mathbf{r}_2\|_2^2\le h} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1 \\ &= \text{Vol}(\|\mathbf{r}_2\|_2^2 \le h)\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\boldsymbol{\xi}_h)d\mathbf{r}_1\end{aligned}$$

where the last equality follows from the mean value theorem with $\boldsymbol{\xi}_h \in B_h^{p-q}(0)$, $B_h^{p-q}(0)$ is the $p-q$ dimensional ball at the origin with radius h.

The numerator of (35) equals

$$\begin{aligned}\int_{\{\mathbf{z}\le\mathbf{x}\}\cap\{\mathbf{z}\in\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}\}} f_{\mathbf{X}}(\mathbf{z})d\mathbf{z} &= \int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q}\int_{\|\mathbf{r}_2\|_2^2\le h} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1 \\ &= \text{Vol}(\|\mathbf{r}_2\|_2^2 \le h)\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\widetilde{\boldsymbol{\xi}}_h)d\mathbf{r}_1\end{aligned}$$

where $(y_1,\dots,y_q)^\top = \mathbf{V}^\top(\mathbf{x}-\mathbf{s}_0)$ and $\widetilde{\boldsymbol{\xi}}_h \in B_h^{p-q}(0)$. Observe that if $\mathbf{x} \notin \mathbf{s}_0+\text{span}\{\mathbf{V}\}$, $(y_1,\dots,y_q)^\top = 0$ and therefore the cdf is constant and the density is 0. Substituting the numerator and denominator into (35) yields

$$\lim_{h\downarrow 0}\frac{\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\widetilde{\boldsymbol{\xi}}_h)d\mathbf{r}_1}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\boldsymbol{\xi}_h)d\mathbf{r}_1} \tag{36}$$

By the dominated convergence theorem, the limit can be passed under the integral, separately for the numerator and denominator since one can choose $M > 0$ such that the integral is negligible outside of $B_M(0)$. On the compact set the continuity of the density obtains an integrable majorant. Since both the numerator and denominator converge, (36) converges to

$$\frac{\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1}$$

Taking the derivative results in (5). Due to the independence of $\mathbf{X}$ and ϵ in (1), $\mathbb{V}\text{ar}(Y|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}) = \mathbb{V}\text{ar}(g(\mathbf{B}^\top\mathbf{X})|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}) + \mathbb{V}\text{ar}(\epsilon)$. Using the density formula in (5) we obtain (7).

The parameter integral Heuser (1995),

$$t^{(l)}(\mathbf{V},\mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}^\top\mathbf{s}_0+\mathbf{B}^\top\mathbf{V}\mathbf{r})^l f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r})d\mathbf{r} = \int_{\mathbb{R}^q}\tilde{g}(\mathbf{V},\mathbf{s}_0,\mathbf{r})d\mathbf{r}$$

is well defined and continuous if (1) $\tilde{g}(\mathbf{V},\mathbf{s}_0,\cdot)$ is integrable for all $\mathbf{V},\mathbf{s}_0$, (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 Heuser (1995) p. 101]. Furthermore $t^{(l)}(\mathbf{V},\mathbf{s}_0) = \int_K \tilde{g}(\mathbf{V},\mathbf{s}_0,\mathbf{r})d\mathbf{r}$ for some compact set K since $\text{supp}(f_{\mathbf{X}})$ is compact. The function $\tilde{g}(\mathbf{V},\mathbf{s}_0,\mathbf{r})$ is continuous in all inputs by the continuity of g and $f_{\mathbf{X}}$, 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 $\boldsymbol{\Sigma}_{\mathbf{X}} > 0$. Then, $\tilde{L}(\mathbf{V},\mathbf{s}_0)$ in (7) is continuous. Since $L(\mathbf{V})$ is a parameter integral, it is well defined and continuous following the same arguments as above. □

Proof of Theorem 8: Since $(\mathbf{X}_i^\top, Y_i)$ are iid draws from the joint distribution of $(\mathbf{X}^\top, Y)$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) &= \frac{1}{nh_n^{p-q}}\mathbb{V}\mathrm{ar}\left(K\left(\frac{d_1(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) Y_1^l\right) \\ &\leq \frac{1}{nh_n^{p-q}}\mathbb{E}\left(K\left(\frac{d_1(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)^2 Y_1^{2l}\right) \leq \frac{\mathbb{E}(Y_1^{2l})M_2^2}{nh_n^{p-q}} \to 0\end{aligned}$$

where the last inequality derives from the boundedness of the kernel, $K(\cdot) \leq M_2$. □

Proof of Theorem 9: Let $\mathbf{U}$ be an orthonormal basis of the orthogonal complement of $\mathrm{span}\{\mathbf{V}\}$ and $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$, where $\mathbf{r}_1 = \mathbf{V}^\top(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^q$, $\mathbf{r}_2 = \mathbf{U}^\top(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^{p-q}$, and $Q_{\mathbf{V}}\mathbf{x} = (\mathbf{I}_p - \mathbf{P}_{\mathbf{V}})\mathbf{x} = \mathbf{U}\mathbf{r}_2$.

$$\begin{aligned}\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}}\sum_{i=1}^n K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) g(\mathbf{B}^\top\mathbf{X}_i)^l\right) &= \frac{1}{h_n^{(p-q)/2}}\mathbb{E}\left(K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) g(\mathbf{B}^\top\mathbf{X}_1)^l\right) \\ &= \frac{1}{h_n^{(p-q)/2}}\int_{\mathbb{R}^p} K\left(\frac{\|Q_{\mathbf{V}}(\mathbf{x} - \mathbf{s}_0)\|_2^2}{h_n}\right) g(\mathbf{B}^\top\mathbf{x})^l f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\ &= \frac{1}{h_n^{(p-q)/2}}\int_{\mathbb{R}^p} K\left(\frac{\|Q_{\mathbf{V}}\mathbf{x}\|_2^2}{h_n}\right) g(\mathbf{B}^\top\mathbf{s}_0 + \mathbf{B}^\top\mathbf{x})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{x})d\mathbf{x} \\ = \frac{1}{h_n^{(p-q)/2}}\int_{\mathbb{R}^q}\int_{\mathbb{R}^{p-q}} K\left(\|\frac{\mathbf{r}_2}{\sqrt{h_n}}\|_2^2\right) g(\mathbf{B}^\top\mathbf{s}_0 + \mathbf{B}^\top\mathbf{V}\mathbf{r}_1 + \mathbf{B}^\top\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1\end{aligned}$$

Applying Fubini's Theorem and substituting $\tilde{\mathbf{r}}_2 = \mathbf{r}_2/\sqrt{h_n}$, $d\mathbf{r}_2 = h_n^{(p-q)/2}d\tilde{\mathbf{r}}_2$ yields

$$\int_{\mathbb{R}^q}\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|_2^2) g(\mathbf{B}^\top\mathbf{s}_0 + \mathbf{B}^\top\mathbf{V}\mathbf{r}_1 + \sqrt{h_n}\mathbf{B}^\top\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \sqrt{h_n}\mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1$$

By Assumption A.3, Y is integrable. Thus, there exists an $M > 0$ such that the integral outside of $B_M^p(0)$ is negligible. Using similar arguments as in the proof of Theorem 1, the limit can be pulled inside the integral and also inside the functions because of the continuity of $g(\cdot)$ and $f_{\mathbf{X}}(\cdot)$, obtaining (26). Eqns. (27) and (28) follow directly from (26) with $l = 0$ from the independence of $\mathbf{X}_i$ and ϵ_i. □

Proof of Theorem 11: Since $L^2(\Omega)$ convergence implies convergence in probability, (a) and (b) follow from (24), Theorem 10 and the continuous mapping theorem. (c) follows from (a) and (b), Theorem 1 and $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) = \bar{y}_2(\mathbf{V}, \mathbf{s}_0) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0)^2$. □

Proof of Theorem 12: By (14) and (6),

$$|L_n(\mathbf{V}) - L(\mathbf{V})| \leq \frac{1}{n}\sum_i |\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)| + \frac{1}{n}\sum_i |\tilde{L}(\mathbf{V}, \mathbf{X}_i) - \mathbb{E}(\tilde{L}(\mathbf{V}, \mathbf{X}))| \tag{37}$$

The second term on the right hand side goes to 0 almost surely by the strong law of large numbers. For the first term

observe that

$$t_n^{(l)}(\mathbf{V},\mathbf{X}_i)|(\mathbf{X}_i=\mathbf{s}_0)=\frac{1}{nh_n^{(p-q)/2}}K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)Y_i^l+\frac{1}{nh_n^{(p-q)/2}}\sum_{j\neq i}K\left(\frac{d_j(\mathbf{V},\mathbf{s}_0)}{h_n}\right)Y_j^l$$
$$\xrightarrow{L^2(\Omega)} t^{(l)}(\mathbf{V},\mathbf{s}_0)$$

by similar arguments as in the proof of Theorems 8 and 9, since the first term of the right hand side converges to 0 by $nh_n^{(p-q)/2}\to\infty$. Therefore, $Z_n(\mathbf{V},\mathbf{s}_0):=\tilde{L}_n(\mathbf{V},\mathbf{X}_i)|(\mathbf{X}_i=\mathbf{s}_0)\longrightarrow\tilde{L}(\mathbf{V},\mathbf{s}_0)$ in probability by the continuous mapping theorem.

Under Assumption A.5, $\tilde{L}_n(\mathbf{V},\mathbf{s}_0)\leq 4M_1^2$, $Z_n(\mathbf{V},\mathbf{s}_0)\leq 4M_1^2$ and $L_n(\mathbf{V},\mathbf{s}_0)\leq 4M_1^2$, so that $Z_n(\mathbf{V},\mathbf{s}_0)$ is uniformly integrable. Therefore, $Z_n(\mathbf{V},\mathbf{s}_0)\xrightarrow{L^2(\Omega)}\tilde{L}(\mathbf{V},\mathbf{s}_0)$, which implies convergence in $L^1(\Omega)$. Let $\tilde{Z}_n(\mathbf{s}_0)=\mathbb{E}|Z_n(\mathbf{V},\mathbf{s}_0)-\tilde{L}(\mathbf{V},\mathbf{s}_0)|$. By Assumption A.5, $\tilde{Z}_n(\mathbf{s}_0)\leq 32M_1^2$. Next,

$$\lim_{n\to\infty}\mathbb{E}\left(\frac{1}{n}\sum_i|\tilde{L}_n(\mathbf{V},\mathbf{X}_i)-\tilde{L}(\mathbf{V},\mathbf{X}_i)|\right)=\lim_{n\to\infty}\mathbb{E}\left(|\tilde{L}_n(\mathbf{V},\mathbf{X}_i)-\tilde{L}(\mathbf{V},\mathbf{X}_i)|\right)$$
$$=\lim_{n\to\infty}\mathbb{E}\left(\mathbb{E}|\tilde{L}_n(\mathbf{V},\mathbf{X}_i)-\tilde{L}(\mathbf{V},\mathbf{X}_i)||\mathbf{X}_i=\mathbf{s}_0\right)=\lim_{n\to\infty}\mathbb{E}\left(\tilde{Z}_n(\mathbf{X})\right) \tag{38}$$

$\tilde{Z}_n(\mathbf{s}_0)\to 0$ for all $\mathbf{s}_0$, so that $\tilde{Z}_n(\mathbf{X})\to 0$ almost surely. By dominated convergence, the limit can be swapped with the expectation in (38) which yields that the limit is 0. Therefore, the first term goes to 0 in $L^1(\Omega)$ and the second almost surely in the right hand side of (37). □

Proof of Theorem 5: From (12) and (13) we have $\tilde{L}_n=\bar{y}_2-\bar{y}_1^2$ where $\bar{y}_l=\sum_i w_i Y_i^l$ for $l=1,2$. We suppress the dependence on $\mathbf{V}$ and $\mathbf{s}_0$ and write $w_i=K_i/\sum_j K_j$. For the Gaussian kernel, $\nabla K_i=(-1/h_n^2)K_i d_i\nabla d_i$ and $\nabla w_i=\left(K_i d_i\nabla d_i(\sum_j K_j)-K_i\sum_j K_j d_j\nabla d_j\right)/(\sum_j K_j)^2$. Then

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

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

Derivation of (30): By Theorem 1, the density, dropping the normalization constant, is

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\mathrm{span}\{\mathbf{V}\}}(\mathbf{x})\propto f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}r_1)\propto\exp\left(-\frac{1}{2}(\mathbf{s}_0+r_1\mathbf{V})^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{s}_0+r_1\mathbf{V})\right)$$
$$\propto\exp\left(-\frac{1}{2}\left(2r_1\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0+r_1^2\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}\right)\right)=\exp\left(-\frac{1}{2\sigma^2}\left(2r_1\sigma^2\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0+r_1^2\right)\right)$$
$$\propto\exp\left(-\frac{1}{2\sigma^2}(r_1-\alpha)^2\right), \tag{39}$$

where the square is completed in (39) with $\sigma^2=1/(\mathbf{V}^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V})$ and $\alpha=-\sigma^2\mathbf{V}^\top\Sigma^{-1}\mathbf{s}_0$. Let $\psi(z)$ be the density of a

standard normal variable. Then,

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}}(\mathbf{x}) = \begin{cases} \frac{1}{\sigma}\psi(\frac{r_1-\alpha}{\sigma}) & \text{if } \mathbf{x}\in\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}, r_1=\mathbf{V}^\top(\mathbf{x}-\mathbf{s}_0)\in\mathbb{R} \\ 0 & \text{otherwise} \end{cases}$$