Conditional Variance Estimator for Sufficient Dimension Reduction

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

June 27, 2019

Abstract

Conditional Variance Estimation (CVE) is a novel sufficient dimension reduction (SDR) method for regressions satisfying $\mathbb{E}(Y|\mathbf{X}) = \mathbb{E}(Y|\mathbf{B}'\mathbf{X})$, where $\mathbf{B}'\mathbf{X}$ 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 set-ups, while it remains largely on par under most others.

Keywords: Regression, SDR, mean subspace, MAVE

[*] lukas.fertl@tuwien.ac.at

[†] The authors gratefully acknowledge the support of the Austrian Science Fund (FWF P 30690-N35).

1 Introduction

Suppose $(Y, \mathbf{X}')'$ 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

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

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

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

and requires the first conditional moment $\mathbb{E}(Y|\mathbf{X}) = g(\mathbf{B}'\mathbf{X})$ contain the entirety of the information in $\mathbf{X}$ about Y to be captured by $\mathbf{B}'\mathbf{X}$, so that $F(Y|\mathbf{X}) = F(Y|\mathbf{B}'\mathbf{X})$, 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 $\mathbf{B}'\mathbf{X}$ is given and replacing $\mathbf{X}$ by $\mathbf{B}'\mathbf{X}$ 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., $\mathbf{B}'\mathbf{B} = \mathbf{I}_k$, since a change of coordinate system by an orthogonal transformation does not alter model (2).

The first split in SDR taxonomy occurs at likelihood versus non-likelihood based methods. The former, which were developed more recently [6, 5, 7, 3, 2], assume knowledge either of the joint family of distributions of $(Y, \mathbf{X}')'$, 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 [1, 17, 14].

The method we propose, the conditional variance estimator, falls in the class of nonparametric methods. The estimators in this class minimize a criterion that describes the fit of the dimension reduction models (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, [14]]. Principal Hessian Directions (pHd, [16]) 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* [8], 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, [23]). MAVE assumes model (1), bounded fourth derivative covariate density, and g having continuous bounded third derivative. It is based on a local first order approximation of g in (1) and the minimization of the expected conditional variance of the response given $\mathbf{B}'\mathbf{X}$.

The conditional variance estimator (CVE) also targets and recovers span$\{\mathbf{B}\}$ in models (1) and (2). The objective function is based on an intuitive idea regarding the directions in the predictor space that capture the dependence of Y on $\mathbf{X}$ along which Y exhibits significantly higher variation in contrast to the orthogonal directions along which Y exhibits markedly less variation. CVE is a fully data-driven estimator that is seen to perform on par with MAVE in normal predictor regressions and to outperform it in regressions with covariate distributions outside the elliptically contoured family, such as normal mixtures, in simulations. CVE is applicable to regressions with $p > n$ as it targets the orthogonal complement of span$\{\mathbf{B}\}$ and, thus, circumvents the inversion of $\mathbf{\Sigma}_{\mathbf{x}}$. 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 organized 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 analyzed 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 distribution and denote its support by $\mathrm{supp}(f_{\mathbf{X}})$. We refer to the following assumptions when needed in the sequel.

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, $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}\mathrm{ar}(\epsilon) = \eta^2 < \infty$, and $\mathbf{\Sigma}_{\mathbf{x}}$ is positive definite.*

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

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

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

is 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 [22] and Section 2.1 of [19]]. For $q \leq p \in \mathbb{N}$ and any $\mathbf{V} \in S(p,q)$, we define

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

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

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

Theorem 1 establishes that (5) is a proper density and that $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (4), and its generalized 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 [11]. Moreover, Theorem (1) provides its explicit 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

$$\mu_l(\mathbf{V}, \mathbf{s}_0) := \int_{\mathbb{R}^q} g(\mathbf{B}'\mathbf{s}_0 + \mathbf{B}'\mathbf{V}\mathbf{r}_1)^l \frac{f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} d\mathbf{r}_1 = \frac{t^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0)}$$

with $t^{(l)}(\mathbf{V}, \mathbf{s}_0) := \int_{\mathbb{R}^q} g(\mathbf{B}'\mathbf{s}_0 + \mathbf{B}'\mathbf{V}\mathbf{r}_1)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1$.

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* $\mathbf{v}_u \in \operatorname{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \mathbb{V}\mathrm{ar}(\epsilon) = \eta^2$.

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

(c) For all $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q)$ *such that there exist* $u \in \{1, ..., q\}$ *with* $\mathbf{v}_u \in \operatorname{span}\{\mathbf{B}\}$, $L(\mathbf{V}) > \eta^2$.

(d) For all $\mathbf{V} \in \operatorname{span}\{\mathbf{B}\}^\perp$, $L(\mathbf{V}) = \eta^2$

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

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

since $\mathbf{B}'\mathbf{V}\mathbf{V}'(\mathbf{X}-\mathbf{s}_0) \neq 0$ w.p. 1, and therefore the first term in (8) has positive variance. For $\mathbf{V}$ such that $\mathbf{V} \perp \mathbf{B}$, $\mathbf{B}'\mathbf{V}\mathbf{V}'(\mathbf{X}-\mathbf{s}_0) = 0$ 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 $Y = g(\mathbf{B}'\mathbf{X})$ 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 $g(\mathbf{B}'\mathbf{X})$. In consequence, the objective function (6) captures the variation of Y as $\mathbf{X}$ varies in the column space of $\mathbf{V}$ and is minimized 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

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

$\mathbf{V}_q$ is well defined as the minimizer 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 optimization 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 minimizer $\mathbf{V}_{p-k}$ is not unique since for all $\mathbf{C} \in \mathbb{R}^{q \times p-k}$ such that $\mathbf{C}\mathbf{C}' = \mathbf{I}_{p-k}$, $L(\mathbf{V}\mathbf{C}) = L(\mathbf{V})$ as $L(\mathbf{V})$ depends on $\mathbf{V}$ only through $\operatorname{span}\{\mathbf{V}\}$. Nevertheless, since every minimizer 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 the Conditional Variance Estimator at the population level.

Definition 4. *Let*

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

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

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 $\mathbf{X} = \mathbf{s}_0 + \mathbf{V}(\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V})^{-1}\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{X} - \mathbf{s}_0)$, 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 $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{y}$. 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 $(Y_i, \mathbf{X}_i')'_{i=1,\dots,n}$ is an i.i.d. sample from model (1). We define

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

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P_V} = \mathbf{VV}'$ and $\mathbf{Q_V} = \mathbf{I}_p - \mathbf{P_V}$. 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. $|K(\cdot)| \leq M_1$) with $\int_{\mathbb{R}^q} K(\|\mathbf{r}\|_2^2)d\mathbf{r} < \infty$ for $q \leq p-1$, which we refer to as *kernel*. Examples of such functions include the rectangular, $K(z) = cI(z \leq 1)$, 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 Table 1 of [18]. For $i = 1, \ldots, n$, 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.

$L_n(\mathbf{V})$ 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 an 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 Lemma 5, which is proven in the Appendix.

Lemma 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 Choosing the bandwidth h_n

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) $\lim_{n \to \infty} h_n = 0$, (b) $\lim_{n \to \infty} n h_n^{p-q} = \infty$, and (c) $\lim_{p-q \to \infty} h_n = \infty$. We use a heuristically motivated rule that performs well in simulation studies.

Lemma 6. *Let* $\mathbf{M}$ *be a* $p \times p$ *positive definite matrix. Then,*

$$\frac{tr(\mathbf{M})}{p} = \text{argmin}_{s>0} \|\mathbf{M} - s\mathbf{I}_p\|_2 \tag{15}$$

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, $\mathbf{M} = \mathbf{U}\text{diag}(\lambda_1, ..., \lambda_p)\mathbf{U}'$, 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 (15), 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 $N(\boldsymbol{\mu}_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$, for $\sigma^2 = \text{tr}(\boldsymbol{\Sigma}_{\mathbf{x}})/p$, by Lemma 6. 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, $d_i(\mathbf{V}, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|_2^2 - (\mathbf{X}_i - \mathbf{X}_j)'\mathbf{V}\mathbf{V}'(\mathbf{X}_i - \mathbf{X}_j) = \|\tilde{\mathbf{X}}_i\|_2^2 - \tilde{\mathbf{X}}_i'\mathbf{V}\mathbf{V}'\tilde{\mathbf{X}}_i$. Let

$$\begin{aligned} \text{nObs} &:= \mathbb{E}\big(\#\{i \in \{1, ..., n\} : \tilde{\mathbf{X}}_i \in \text{span}_h\{\mathbf{V}\}\}\big) \\ &= 1 + (n-1)\mathbf{P}(d_1(\mathbf{V}, \mathbf{X}_2) \leq h) = 1 + (n-1)\mathbf{P}(\|\tilde{\mathbf{X}}\|_2^2 - \tilde{\mathbf{X}}'\mathbf{V}\mathbf{V}'\tilde{\mathbf{X}} \leq h) \end{aligned} \tag{16}$$

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 (16).

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

$$\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2, \tag{17}$$

where $\mathbf{r}_1 = \mathbf{V}'\mathbf{x}$, $\mathbf{r}_2 = \mathbf{U}'\mathbf{x}$ and $\mathbf{U}'\mathbf{V} = \mathbf{0}, \mathbf{U}'\mathbf{U} = \mathbf{I}_{p-q}$. By (17), $\tilde{\mathbf{X}} = \mathbf{V}\mathbf{R}_1 + \mathbf{U}\mathbf{R}_2$, with $\mathbf{R}_1 = \mathbf{V}'\tilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_q), \mathbf{R}_2 = \mathbf{U}'\tilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_{p-q})$. Then, $\tilde{\mathbf{X}}'\mathbf{V}\mathbf{V}'\tilde{\mathbf{X}} = \|\mathbf{R}_1\|_2^2$ and

$\|\tilde{\mathbf{X}}\|_2^2 = \|\mathbf{R}_1\|_2^2 + \|\mathbf{R}_2\|_2^2$. Therefore,

$$\mathbf{P}(\|\tilde{\mathbf{X}}\|_2^2 - \tilde{\mathbf{X}}'\mathbf{V}\mathbf{V}'\tilde{\mathbf{X}} \leq h) = \mathbf{P}(\|\mathbf{R}_2\|_2^2 \leq h) = \chi_{p-q}\left(\frac{h}{2\sigma^2}\right), \tag{18}$$

where χ_{p-q} is the cdf of a chi-squared distribution with $p-q$ degrees of freedom. Plugging (18) in (16) obtains

$$\text{nObs} = 1 + (n-1)\chi_{p-q}\left(\frac{h}{2\sigma^2}\right). \tag{19}$$

Solving (19) for h and Lemma 6 yield the bandwidth used in the simulation studies,

$$h_n(\text{nObs}) := \chi_{p-q}^{-1}\left(\frac{\text{nObs}-1}{n-1}\right)\frac{2\text{tr}(\widehat{\mathbf{\Sigma}}_{\mathbf{x}})}{p}, \tag{20}$$

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

In order to ascertain h_n satisfies conditions (a), (b) and (c) 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 $\gamma(n)/n \to 0$. In the simulations in Section 7, nObs $= \gamma(n) = n^\beta$ with $\beta \in (0,1)$ is used.

Since the ad-hoc procedure described in this section yields satisfactory results, we opted against cross validation because of the computational burden involved. Specifically, given h_n, one would estimate $\mathbf{B}$, fit a forward model with the projected data, apply cross validation and select h_n that obtains the lowest cross-validated error in the forward model.

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 (7) a curvilinear search is used [24, 19], 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 [4]. A step-size $\tau > 0$ and tolerance $\text{tol} > 0$ are fixed at the outset.

```
Result: $\mathbf{V}^{(\text{end})}$
Initialize: $\mathbf{V}^{(0)}$, $\tau = 1$, $\text{tol} = 10^{-3}$, $\text{error} = \text{tol} + 1$, $\text{maxit} = 50$, $\text{count} = 0$;
while $error > tol$ and $count \leq maxit$ do
    • $\mathbf{G} = \nabla_{\mathbf{V}} L_n(\mathbf{V}^{(j)}) \in \mathbb{R}^{p \times q}$, $\mathbf{W} = \mathbf{G}\mathbf{V}' - \mathbf{V}\mathbf{G}' \in \mathbb{R}^{p \times p}$
    • $\mathbf{V}^{(j+1)} = (\mathbf{I}_p + \tau \mathbf{W})^{-1}(\mathbf{I}_p - \tau \mathbf{W})\mathbf{V}^{(j)}$
    • $\text{error} = \|\mathbf{V}^{(j)}\mathbf{V}^{(j')} - \mathbf{V}^{(j+1)}\mathbf{V}^{(j+1)'}\|_2/(pq)$
    if $L_n(\mathbf{V}^{(j+1)}) - L_n(\mathbf{V}^{(j)}) > 0$ then
        $\mathbf{V}^{(j+1)} \leftarrow \mathbf{V}^{(j)}$; $\tau \leftarrow \frac{\tau}{2}$; $\text{error} \leftarrow \text{tol} + 1$
    else
        $\text{count} \leftarrow \text{count} + 1$
    end
end
```

Algorithm 1: Curvilinear search

[24] showed that the sequence generated by the algorithm converges to a stationary point if Armijo-Wolfe conditions are used for determining the stepsize τ. We opted for

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

$$\bar{y}_l(\mathbf{V},\mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V},\mathbf{s}_0)}{t_n^{(0)}(\mathbf{V},\mathbf{s}_0)}, \tag{21}$$

where

$$t_n^{(l)}(\mathbf{V},\mathbf{s}_0) = \frac{1}{nh_n^{(p-q)/2}}\sum_{i=1}^n K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)Y_i^l \tag{22}$$

for $l = 0,1,2$.

Theorem 7. *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* (22), $l = 0,1,2$.

Theorem 8. *Under assumptions A.1, A.2, A.4, and $h_n \to 0$,*

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}}\sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right) g(\mathbf{B}'\mathbf{X}_i)^l\right) \to t^{(l)}(\mathbf{V},\mathbf{s}_0)\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}\|_2^2)d\mathbf{r}, \tag{23}$$

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}}\sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\epsilon_i\right) = 0, \tag{24}$$

and

$$\mathbb{E}\left(\frac{1}{nh_n^{(p-q)/2}}\sum_{i=1}^{n} K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\epsilon_i^2\right) \to \eta^2 t^{(0)}(\mathbf{V},\mathbf{s}_0)\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}\|_2^2)d\mathbf{r} \tag{25}$$

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

Theorem 9. *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 (22) and $t^{(l)}$ defined in Theorem 1, for $l = 0, 1, 2$.

Theorem 9 follows directly from Theorems 7, 8 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 10. *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* (21) *and* $\mu_l(\cdot, \cdot)$ *in Theorem 1 for* $l = 1, 2$.

Theorems 7-10 lead to Theorem 11 that establishes the consistency of the sample CVE objective function.

Theorem 11. *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 $L_n(\mathbf{V})$ behavior

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

Applying Theorem 1 obtains

$$\mu_l(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^2} g(\mathbf{B}'\mathbf{x})^l f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^2} (\mathbf{B}'\mathbf{x})^l f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) d\mathbf{x} \tag{26}$$

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

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x}) = \begin{cases} \frac{1}{\sigma}\psi(\frac{r_1-\alpha}{\sigma}) & \text{if } \mathbf{x} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}, r_1 = \mathbf{V}'(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R} \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

where $\psi(z)$ is the density of a standard normal variable. Inserting (27) in (26), we obtain

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

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

$$L(\mathbf{V}) = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right) = (\mathbf{B}'\mathbf{V})^2\sigma^2 + \eta^2 = \frac{(\mathbf{B}'\mathbf{V})^2}{\mathbf{V}'\mathbf{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}} + \eta^2 \tag{28}$$

From (28) we clearly see that $L(\mathbf{V})$ attains its minimum when $\mathbf{V} \perp \mathbf{B}$. Also, if $\mathbf{\Sigma}_{\mathbf{x}} = \mathbf{I}_2$, 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 $\mathbf{V}(\theta) = (\cos(\theta), \sin(\theta))'$, $\theta \in [0, \pi]$. Since $\mathbf{B} = (1, 0)'$, the minimum of $\tilde{L}(\mathbf{V})$ is at $\mathbf{V}(\pi/2) = (0, 1)' \perp \mathbf{B}$.

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.

6 Simulation studies

We compare the estimation accuracy of CVE with the forward model based SDR methods, mean MAVE (`meanMAVE`) [23], central subspace MAVE (`csMAVE`) [20] and `pHd` [16, 10], and the inverse regression based methods, SIR [15] and SAVE [9]. Central subspace MAVE (`csMAVE`) assumes $Y = g(\mathbf{B}'\mathbf{X}, \epsilon)$, which is a much more general model than (1). `csMAVE`

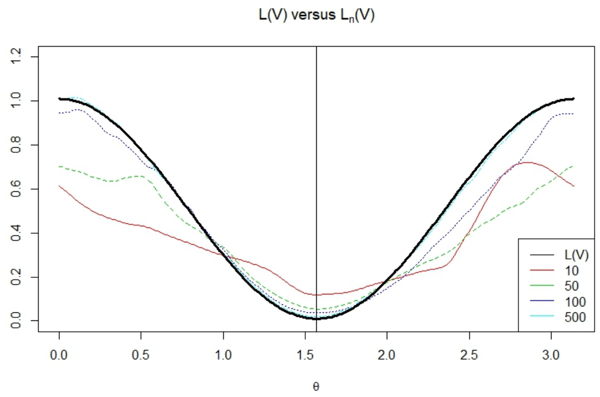

Figure 1: Solid black: $L(\mathbf{V}(\theta)) = \cos(\theta)^2 + 0.1^2$, colored $L_n(\mathbf{V}(\theta))$, $\theta \in [0, \pi]$, $n = 10, 50, 100, 500$. Vertical black line at $\theta = \pi/2$

is included in the comparison as it is one of only three existing forward model based SDR methods. The dimension k is assumed to be known throughout.

We implement CVE using 30 arbitrary starting values in the optimization algorithm. We use three different bandwidths, $h_n(n^{0.8}) > h_n(n^{0.66}) > h_n(n^{0.5})$ in (20), which are indicated with CVE1, CVE2 and CVE3, respectively. Except for CVE, these methods are implemented using the `R` packages `dr` and `MAVE`.

The five models we consider are given in Table 1. Throughout, we set $p = 12$, $\mathbf{b}_1 = (1,1,1,1,1,1,0,0,0,0,0,0)'/\sqrt{6}$, $\mathbf{b}_2 = (1,-1,1,-1,1,-1,0,0,0,0,0,0)'/\sqrt{6}$, except for M2 where $\mathbf{b}_1 = \mathbf{e}_1$ and $\mathbf{b}_2 = \mathbf{e}_2$, where $\mathbf{e}_j$ denotes the j unit vector in $\mathbb{R}^{12}$, and ϵ is standard normal independent of $\mathbf{X}$.

The variance-covariance structure of $\mathbf{X}$ in models M1 and M3 satisfies $\mathbf{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $i, j = 1, \ldots, p$. M1 is studied in both [23] and [15], but we use $p = 12$ instead of 10 and a non identity covariance structure. In M4, $Z \sim$ Bernoulli(p_{mix}), where $\mathbf{1}_q = (1, 1, ..., 1)' \in \mathbb{R}^q$, mixing probability $p_{\text{mix}} \in [0, 1]$ and dispersion parameter $\lambda > 0$. For $0 < p_{\text{mix}} < 1$, $\mathbf{X}$ has

Table 1: Models

Name	Model	$\mathbf{X}$ distribution	k	n
M1	$Y = \frac{\mathbf{b}_1'\mathbf{X}}{0.5+(\mathbf{b}_2'\mathbf{X}+1.5)^2} + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	2	200
M2	$Y = (\mathbf{b}_1'\mathbf{X})(\mathbf{b}_2'\mathbf{X})^2 + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{I}_p)$	2	200
M3	$Y = \cos(\mathbf{b}_1'\mathbf{X}) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	1	100
M4	$Y = \cos(\mathbf{b}_1'\mathbf{X}) + 0.5\epsilon$	$\mathbf{X} \sim (Z(-\mathbf{1}_{12}) + (1-Z)\mathbf{1}_{12})\,\lambda + N_p(0, \mathbf{I}_p)$	1	100
M5	$Y = 2\log(\lvert\mathbf{b}_1'\mathbf{X}\rvert + 1) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{I}_p)$	1	42

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-M5. We estimate $\mathbf{B}$ using the six 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 $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}\mathbf{B}')^{-1}\mathbf{B}'$ is the orthogonal projection matrix on span$\{\mathbf{B}\}$. The factor $\sqrt{2k}$ normalizes the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement.

In Figures 2 - 4 we plot the box-plots of the $r = 100$ estimation errors for each method. For models M1-M3, CVE is approximately on par with MAVE, its main competitor, as can be seen in Figs 2 - 3. SIR and SAVE are not competitive throughout our experiments. SIR, in particular, is expected to fail in models M3-M5 since $\mathbb{E}(Y|\mathbf{X})$ is even.

CVE shows its advantage in Figure 3, where box-plots of the errors in models M3 and M5 are plotted, and Figure 4 with the box-plots for model M4. Model M5 depicts the setting where the sample size is small (42) relative to the predictor dimension (12). The value of the sample size was selected so that MAVE applies without "pre-screeening," which carries out some form of model selection that is unspecified in the `MAVE` package documentation. In Fig. 4, 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. 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 `meanMAVE` and `csMAVE` 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. Since models M3 and M4 are otherwise identical, CVE appears to have an advantage over MAVE when the predictors have mixture distributions. CVE is the only method that estimates the mean subspace reliably in model M4 (err $\approx$ 0.3 to 0.4), whereas MAVE misses it completely (err $\approx$ 1). These results indicate that CVE is either approximately on par, or can perform better than MAVE depending on the predictor distribution.

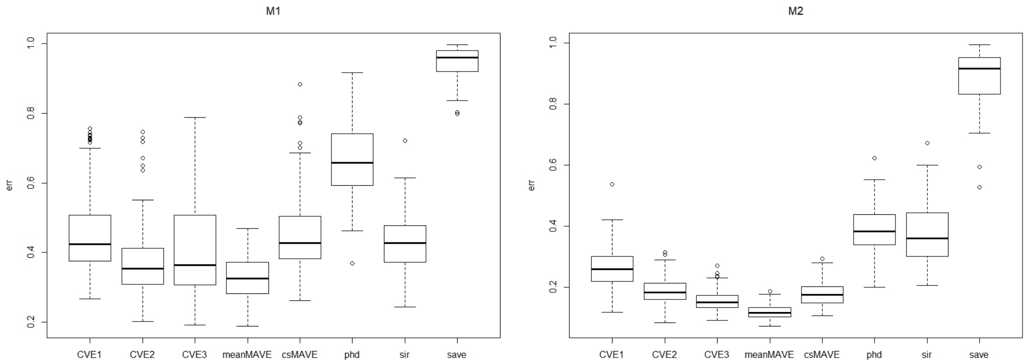

Figure 2: Left panel: M1, $p = 12, n = 200$; Right panel: M2, $p = 12, n = 200$

7 Hitters data set

The Hitters data were analyzed by [23]. The response is $Y = \log(\text{salary})$ and the covariate vector is the 16-dimensional $\mathbf{X} = (x_1, ..., x_{16})'$. 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,

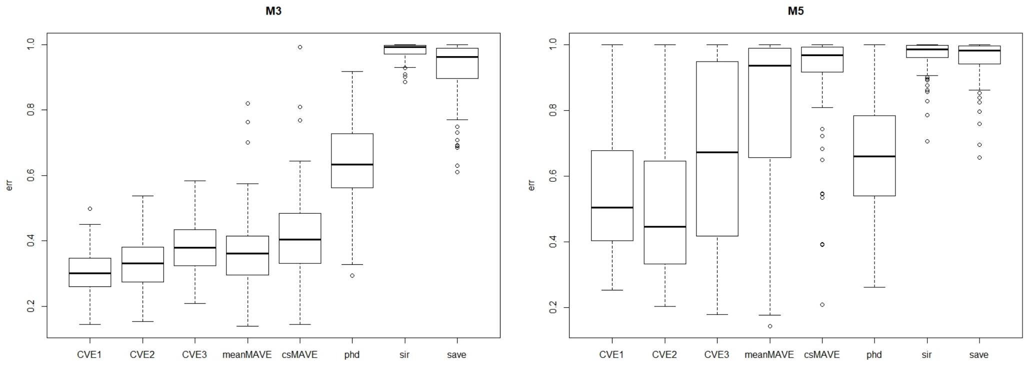

Figure 3: Left panel: M3, $p = 12, n = 100$; Right panel: M5 $p = 12, n = 42$

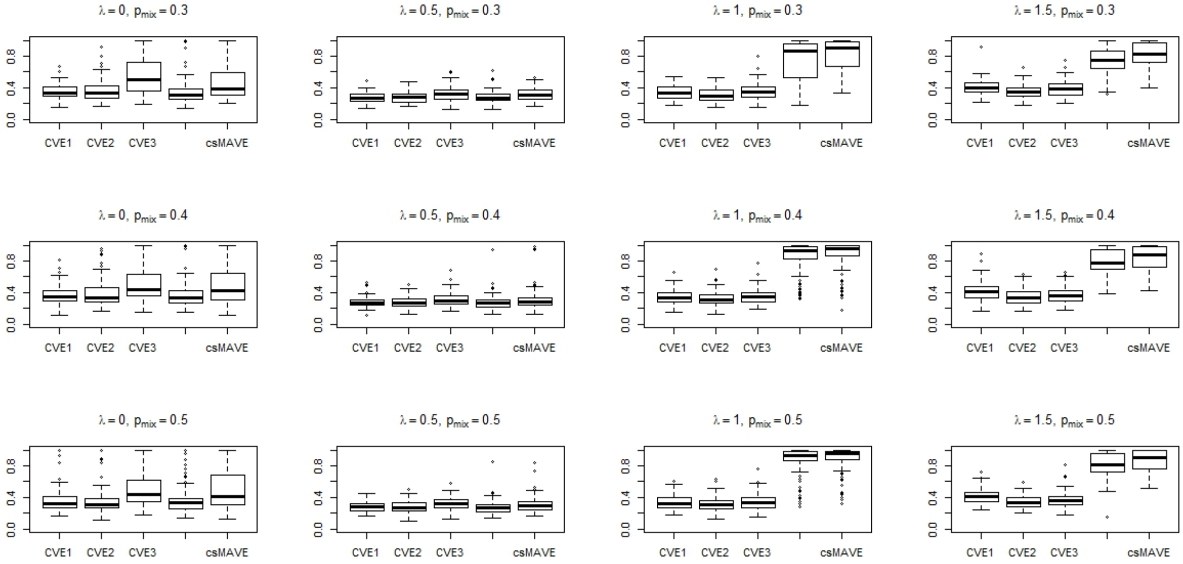

Figure 4: M4, $p = 12, n = 100$

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 [23], we standardize $\mathbf{X}$ 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 [23].

We estimate the dimension k via cross-validation, following the approach in [23], with

$$\hat{k} = \mathrm{argmin}_{l=0,\dots,p} CV(l), \tag{29}$$

where $CV(l) = \sum_i (Y_i - \hat{g}^{-i}(\widehat{\mathbf{B}}_l' \mathbf{X}_i))^2/n$, $\hat{g}^{-i}(\cdot) = \sum_{j=1, j\neq i}^n \tilde{w}_j(\cdot) Y_j$ is the local linear smoother [12, 21], $CV(0) = \sum_i (Y_i - \bar{Y})^2/n$ with $\bar{Y} = \sum_i Y_i/n$, and $\widehat{\mathbf{B}}_l = \widehat{\mathbf{V}}_{p-l}^{\perp}$ is any basis of the orthogonal complement of $\widehat{\mathbf{V}}_{p-l}$, with

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

For a given l, we calculate $\widehat{\mathbf{B}}_l$ from the whole data set and predict Y_i by $\hat{Y}_{i,l} = \hat{g}^{-i}(\widehat{\mathbf{B}}_l' \mathbf{X}_i) = \sum_{j=1, j\neq i}^n \tilde{w}_j(\widehat{\mathbf{B}}_l' \mathbf{X}_i / \tilde{h}_{n,l}) Y_j$, using the bandwidth $\tilde{h}_{n,l} = n^{-1/(3+2l)}$. For $l = p$, $\widehat{\mathbf{B}}_p = \mathbf{I}_p$. We set $\mathrm{SqDev}_{i,l} = (Y_i - \hat{Y}_{i,l})^2$.

For CVE, we use four different choices of nObs for the bandwidth. CVE1-3 are as in Section 6 and CVE4 uses $h_n(n^{0.4})$. Table 2 reports the average and median $\mathrm{SqDev}_{i,l}$ over $l = 1, \dots, 5$; i.e., $\sum_i \mathrm{SqDev}_{i,l}/n$ in the first and the median in the second line for each l.

With respect to mean squared deviations, MAVE, which is `meanMAVE` in this application, appears to outperform CVE. It estimates the dimension to be 2, as do CVE1-3, but CVE4 estimates it to be 5. If the median square deviation is used instead to estimate the dimension, all CVE methods would estimate it to be 2, in agreement with MAVE. Inspection of the summary statistics in Table 3 reveals that CVE is distorted by few extremely

Table 2: Mean and Median SqDev

l	CVE1	CVE2	CVE3	CVE4	MAVE
1	0.274	0.411	0.383	0.341	0.428
	0.080	0.056	0.064	0.071	0.090
2	0.244	0.308	0.247	0.268	0.127
	0.076	0.058	0.046	0.040	0.051
3	0.284	0.460	0.268	0.271	0.184
	0.082	0.075	0.058	0.057	0.091
4	0.470	0.350	0.440	0.313	0.260
	0.117	0.093	0.074	0.066	0.154
5	0.588	0.370	0.397	0.261	0.190
	0.148	0.115	0.098	0.078	0.083

Table 3: Summary statistics for l=2

$l = 2$	CVE1	CVE2	CVE3	CVE4	MAVE
min	0.000	0.000	0.000	0.000	0.000
Q_1	0.016	0.0147	0.006	0.008	0.014
Median	0.076	0.058	0.046	0.040	0.051
Mean	0.244	0.308	0.247	0.268	0.127
Q_3	0.257	0.187	0.177	0.188	0.188
max	13.971	33.544	15.014	9.901	0.998

high squared deviations.

Since CVE4 has the lowest median and maximum, it is used for further analysis. Following [23], we plot the response against the estimated directions in Fig. 5. CVE and MAVE pick up the same pattern: the response appears to be linear in one direction and quadratic in the second.

For CVE4, the fitted regression is

$$\hat{Y} = 0.360915 + 0.269121(b_1'\mathbf{X}) + 0.345169(b_2'\mathbf{X}) - 0.071651(b_2'\mathbf{X})^2 \tag{30}$$

with $R^2 = 0.7729$, and for MAVE

$$\hat{Y} = 0.39051 + 0.49546(b_1'\mathbf{X}) + 1.32529(b_2'\mathbf{X}) - 0.55328(b_2'\mathbf{X})^2 \tag{31}$$

with $R^2 = 0.7859$. Both models (30) and (31) have about the same fit as measured by R^2.

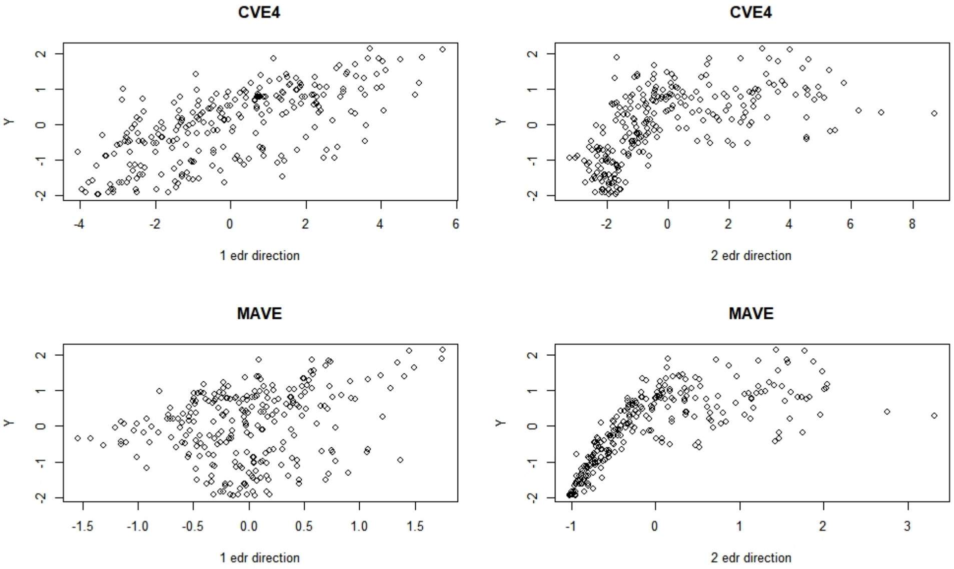

Figure 5: Y against $b_1'\mathbf{X}$ and $b_2'\mathbf{X}$

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.

The theoretical challenge that CVE presents arises from the novelty of its definition that involves random weights that depend on the parameter to be estimated. This precludes the usage of standard probabilistic arguments for establishing consistency of the subspace estimates and may require new probability tools.

CVE does not involve the estimation of the link function g in (1), which may explain why CVE has an advantage over mean MAVE, its direct competitor, in some regression settings. Moreover, CVE does not require the inversion of the predictor covariance matrix and can be applied to regressions with $p \approx n$ or $p > n$.

CVE is a nonparametric estimation technique and the bandwidth choice is important. Our results were obtained by a heuristic rule based on a reparametrization. With this choice of bandwidth, CVE is shown to exhibit similar or better estimation performance than MAVE in our simulations.

Improvement of CVE performance via bias reduction techniques, a complete study of its asymptotic properties, optimal bandwidth selection and extension to central space estimation are under investigation.

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] 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.

[3] 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.

[4] Yasuko Chikuse. *Invariant measures on Stiefel manifolds with applications to multi-*

variate analysis, volume Volume 24 of *Lecture Notes–Monograph Series*, pages 177–193. Institute of Mathematical Statistics, Hayward, CA, 1994.

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

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

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

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

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

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

[11] D. Leao Jr. et al. Regular conditional probability, disintegration of probability and Radon spaces. *Proyecciones*, 23(1):15–29, 05 2004.

[12] Jianqing Fan and Irène Gijbels. *Local polynomial modelling and its applications.* Number 66 in Monographs on statistics and applied probability series. Chapman & Hall, London [u.a.], 1996.

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

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

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

[16] 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.

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

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

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

[20] Hansheng Wang and Yingcun Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821, 2008.

[21] L Wasserman. *All of nonparametric statistics*. Springer, New York, 2006.

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

[23] 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.

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

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

$$\mathbf{P}(\mathbf{X} \leq \mathbf{x}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \lim_{h \downarrow 0} \frac{\mathbf{P}(\{\mathbf{X} \leq \mathbf{x}\} \cap \{\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}\})}{\mathbf{P}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\})} \tag{32}$$

where $\text{span}_h\{\mathbf{V}\}$ is defined below (16). Let $\mathbf{U}$ be an orthonormal basis of the orthogonal complement of $\text{span}\{\mathbf{V}\}$; that is, $\mathbf{U}'\mathbf{V} = \mathbf{0}, \mathbf{U}'\mathbf{U} = \mathbf{I}_{p-q}$. Let $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$ where $\mathbf{r}_1 = \mathbf{V}'(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^q, \mathbf{r}_2 = \mathbf{U}'(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^{p-q}$. Then,

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

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

The numerator of (32) equals

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

where $(y_1, ..., y_q)' = \mathbf{V}'(\mathbf{x} - \mathbf{s}_0)$ and $\widetilde{\boldsymbol{\xi}}_h \in B_h^{p-q}(\mathbf{0})$. Observe that if $\mathbf{x} \notin \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, $(y_1, ..., y_q)' = 0$ and therefore the cdf is constant and the density is 0. Substituting the numerator and denominator into (32) yields

$$\lim_{h \downarrow 0} \frac{\int_{-\infty}^{y_1} ... \int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\widetilde{\boldsymbol{\xi}}_h) d\mathbf{r}_1}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\boldsymbol{\xi}_h) d\mathbf{r}_1} \tag{33}$$

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 $B_M(\mathbf{0})$. On the compact set the continuity of the density obtains an integrable majorant. Since both the numerator and denominator converge, (33) converges to

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

Taking the derivative results in (5).

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

The parameter integral [13],

$$t^{(l)}(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}'\mathbf{s}_0 + \mathbf{B}'\mathbf{V}\mathbf{r})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r} = \int_{\mathbb{R}^q} \tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) d\mathbf{r}$$

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 [13] 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 $\mathbf{s}_0 \in \operatorname{supp}(f_{\mathbf{X}})$ 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 7: Since $(\mathbf{X}_i', Y_i)$ are iid draws from the joint distribution of $(\mathbf{X}', Y)$,

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

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

Proof of Theorem 8: Let $\mathbf{U}$ be an orthonormal basis of the orthogonal complement of $\operatorname{span}\{\mathbf{V}\}$ and $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$, where $\mathbf{r}_1 = \mathbf{V}'(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^q, \mathbf{r}_2 = \mathbf{U}'(\mathbf{x} - \mathbf{s}_0) \in \mathbb{R}^{p-q}$, and $Q_{\mathbf{V}}\mathbf{x} = (\mathbf{I}_p - \mathbf{P}_{\mathbf{V}})\mathbf{x} = \mathbf{U}\mathbf{r}_2$.

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

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

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

By Assumption A.3, Y is integrable. Thus, there exists an $M > 0$ such that the integral outside of $B_M^p(\mathbf{0})$ 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 (23). Eqns. (24) and (25) follow directly from (23) with $l = 0$ from the independence of $\mathbf{X}_i$ and ϵ_i. □

Proof of Theorem 10: Since $L^2(\Omega)$ convergence implies convergence in probability, (a) and (b) follow from (21), Theorem 9 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 11: By (14) and (6),

$$|L_n(\mathbf{V}) - L(\mathbf{V})| \leq \frac{1}{n}\sum_i |\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)| + \frac{1}{n}\sum_i |\tilde{L}(\mathbf{V}, \mathbf{X}_i) - \mathbb{E}(\tilde{L}(\mathbf{V}, \mathbf{X}))| \quad (34)$$

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

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

by similar arguments as in the proof of Theorems 7 and 8, 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, $Z_n(\mathbf{V}, \mathbf{s}_0) \overset{L^2(\Omega)}{\longrightarrow} \tilde{L}(\mathbf{V}, \mathbf{s}_0)$, 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,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left(\frac{1}{n}\sum_i |\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)|\right) &= \lim_{n\to\infty} \mathbb{E}\left(|\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)|\right) \\ = \lim_{n\to\infty} \mathbb{E}\left(\mathbb{E}|\tilde{L}_n(\mathbf{V}, \mathbf{X}_i) - \tilde{L}(\mathbf{V}, \mathbf{X}_i)||\mathbf{X}_i = \mathbf{s}_0\right) &= \lim_{n\to\infty} \mathbb{E}\left(\tilde{Z}_n(\mathbf{X})\right) \qquad (35)\end{aligned}$$

$\tilde{Z}_n(\mathbf{s}_0) \to 0$ 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 (35) 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 (34). □

Proof of Lemma 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

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

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 (27): By Theorem 1, the density, dropping the normalization constant, is

$$\begin{aligned} f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\mathrm{span}\{\mathbf{V}\}}(\mathbf{x}) &\propto f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}r_1) \propto \exp\left(-\frac{1}{2}(\mathbf{s}_0 + r_1\mathbf{V})'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\mathbf{s}_0 + r_1\mathbf{V})\right) \\ \propto \exp\left(-\frac{1}{2}\left(2r_1\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0 + r_1^2\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}\right)\right) &= \exp\left(-\frac{1}{2\sigma^2}\left(2r_1\sigma^2\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{s}_0 + r_1^2\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(r_1 - \alpha)^2\right), \end{aligned} \tag{36}$$

where the square is completed in (36) with $\sigma^2 = 1/(\mathbf{V}'\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V})$ and $\alpha = -\sigma^2\mathbf{V}'\Sigma^{-1}\mathbf{s}_0$. Let $\psi(z)$ be the density of a standard normal variable. Then,

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