Conditional Variance Estimator for Sufficient Dimension Reduction

LUKAS FERTL\textsuperscript{1} and EFSTATHIA BURA\textsuperscript{2,*}

\textsuperscript{1}Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
TU Wien, Vienna, Austria, E-mail: lukas.fertl@tuwien.ac.at

\textsuperscript{2}Institute of Statistics and Mathematical Methods in Economics
Faculty of Mathematics and Geoinformation
TU Wien, Vienna, Austria, E-mail: *efstathia.bura@tuwien.ac.at

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

Keywords: Regression; Nonparametric; Mean subspace; Minimum average variance estimation; Dimension reduction

1. Introduction

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

\[ Y = g(B^T X) + \epsilon, \]  

(1)

where $X$ is independent of $\epsilon$ with positive definite variance-covariance matrix, $\text{Var}(X) = \Sigma_X$, $\epsilon \in \mathbb{R}$ is a mean zero random variable with finite $\text{Var}(\epsilon) = \mathbb{E}(\epsilon^2) = \eta^2$, $g$ is an unknown continuous non-constant function, and $B = (b_1, \ldots, b_k) \in \mathbb{R}^{p \times k}$ of rank $k \leq p$. Model (1) states that

\[ \mathbb{E}(Y \mid X) = \mathbb{E}(Y \mid B^T X) \]  

(2)

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

Identifying the column space or span of $B$, as only the span$\{B\}$ is identifiable, suffices in order to identify the sufficient reduction of $X$ for the regression of $Y$ on $X$. We assume, without loss of generality, that $B$ is semi-orthogonal, i.e., $B^T B = I_k$, since a change of coordinate system by an orthogonal transformation does not alter model (2).
For $q \leq p$, let
\[
S(p,q) = \{V \in \mathbb{R}^{p \times q} : V^TV = I_q\},
\]
denote the Stiefel manifold that comprises of all $p \times q$ matrices with orthonormal columns. This manifold is compact with $\dim(S(p,q)) = pq - q(q+1)/2$ [see [5, 22]]. The set of all $q$-dimensional subspaces in $\mathbb{R}^p$ is the Grassmann manifold,
\[
Gr(p,q) = S(p,q)/S(q,q).
\]
It equals the quotient space of $S(p,q)$ with all $q \times q$ orthonormal matrices $S(q,q)$, and we can see from (4) that only $\text{span}\{B\}$ is identifiable. The goal of sufficient dimension reduction in model (1) is to find a subspace $M \in Gr(p,k)$ such that any basis $B \in S(p,k)$ of $M$ fulfills (1), or, equivalently (2).

Replacing predictors with sufficient reductions in regression and classification without loss of information is called sufficient dimension reduction [10]. The first split in sufficient dimension reduction taxonomy occurs between likelihood and non-likelihood based methods. The former, which were developed more recently [11, 12, 13, 7, 6], assume knowledge either of the joint family of distributions of $(Y, X^T)^T$, or of the conditional family of distributions for $X \mid Y$. The most researched branch of sufficient dimension reduction is non-likelihood based and contains three classes of methods: Inverse regression based, semi-parametric and nonparametric. For a review, see, e.g., [2, 29, 26].

In this paper we present the conditional variance estimation, which falls in the class of nonparametric methods. The estimators in this class minimize a criterion that describes the fit of the dimension reduction model (2) under (1) to the observed data. Since the criterion involves unknown distributions or regression functions, nonparametric estimation is used to recover $\text{span}\{B\}$. Statistical approaches to identify $B$ in (2) include ordinary least squares and nonparametric multiple index models [36]. The least squares estimator, $\Sigma^{-1}_X \text{cov}(X, Y)$, always falls in $\text{span}\{B\}$ [26, Th. 8.3] if $E(X \mid B^T X)$ is linear in the conditioning argument for all $B$. Principal Hessian Directions [28] was the first sufficient dimension reduction estimator to target $\text{span}\{B\}$ in (2). Its main disadvantage is that it requires the so called linearity and constant variance conditions on the marginal distribution of $X$. Its relaxation, Iterative Hessian Transformation [15], still requires the linearity condition in order to recover vectors in $\text{span}\{B\}$.

The most competitive nonparametric sufficient dimension reduction method up to now has been minimum average variance estimation (MAVE, [37]). It assumes model (1), bounded fourth derivative covariate density, and existence of continuous bounded third derivatives for $g$. It uses a local first order approximation of $g$ in (1) and minimizes the expected conditional variance of the response given $B^T X$.

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

2. Motivation

To motivate the development of the conditional variance estimator (CVE) we start this section by a simple example. We consider a bivariate standard normal predictor vector, $X = (X_1, X_2)^T \sim N(0, I_2)$,
and generate the response from \( Y = g(B^T X) + \epsilon = X_1 + \epsilon \), with \( \epsilon \sim N(0, \eta^2) \) independent of \( X \) with \( \eta = 0.1 \). In this setting, \( k = 1 \), \( g(z) = z \in \mathbb{R} \) and \( B = (1, 0)^T \) in model (1) is aligned with the first coordinate axis.

First we draw a sample of size \( n = 100 \) and plot the \( X_i, i = 1, \ldots, n \) in Figure 1, where the color of the points is determined by their corresponding \( Y_i \) values, with small \( Y_i \) values in blue and large in red and the intensity of the color corresponding to their absolute magnitude. In the direction of \( B \), i.e. the \( x \)-axis, the color has high variation, whereas in the orthogonal direction \((0, 1)\), i.e. the \( y \)-axis, the color has low variation arising solely from the error term \( \epsilon \). It is easier to detect patterns in directions of low variability. The estimator we propose aims to identify \( B \) through its orthogonal complement by finding the directions in which the response \( Y \) varies the least as the predictors \( X \) range over an affine subspace. We formalise this idea next.

**Figure 1.** Plot of the \( X_i \) samples from toy model \( Y = X_1 + \epsilon \) with \( n = 100 \), the color of the points are determined by their corresponding \( Y_i \) values, i.e. the low \( Y_i \) values are assigned blue and the higher the \( Y_i \) value the more red the points are. For the left panel \( V = B = (1, 0)^T \), for the right panel \( V = (0, 1) \perp B \), both with shift point \( s_0 = (0, 0)^T \) denoted as black \( \times \). The subspace \( s_0 + \text{span}(V) \) is indicated via the black arrow and the black dotted lines represent the slice \( \{x \in \mathbb{R}^p : \|x - P_{s_0 + \text{span}(V)} x\|^2 \leq h_n\} \).

Let \((\Omega, \mathcal{F}, \mathbb{P})\) be a probability space, and \( X : \Omega \to \mathbb{R}^p \) be a random vector with a continuous probability density function \( f_X \) and denote its support by \( \text{supp}(f_X) \). Throughout, \( \| \cdot \| \) denotes the Frobenius norm for matrices, Euclidean norm for vectors, and scalar product refers to the euclidean scalar product. For any matrix \( M \in \mathbb{R}^{p \times q} \), or linear subspace \( M \), we denote by \( P_M \) the projection matrix on the column space of the matrix or on the subspace, i.e. \( P_M = M(M^T M)^{-1} M^T \in \mathbb{R}^{p \times p} \). For any
For an integer $q$ with $q \leq p$ and any $V \in S(p, q)$, we define

$$L(V, s_0) = \text{Var}(Y \mid X \in s_0 + \text{span}\{V\}),$$

where $s_0 \in \text{supp}(f_X)$ is a shifting point.

**Definition 2.** For $V \in S(p, q)$, we define the objective function,

$$L(V) = \int_{\mathbb{R}^p} L(V, x) f_X(x) dx = \mathbb{E} \left( \tilde{L}(V, X) \right).$$

We derive that both population based functions (6) and (7) are well defined, next. Let $X$ be a $p$-dimensional continuous random vector with density $f_X(x)$, $s_0 \in \text{supp}(f_X) \subset \mathbb{R}^p$, and $V$ belongs to the Stiefel manifold $S(p, q)$ defined in (3). The function

$$f_X|_{X \in s_0 + \text{span}\{V\}}(r_1) = \frac{f_X(s_0 + Vr_1)}{\int_{\mathbb{R}^q} f_X(s_0 + Vr) dr}$$

(A.1). \( Y = g(B^T X) + \epsilon \) with $Y \in \mathbb{R}$, $g : \mathbb{R}^k \to \mathbb{R}$ non constant in all arguments, $B = (b_1, \ldots, b_k) \in \mathbb{R}^{p \times k}$ of rank $k \leq p$, $X \in \mathbb{R}^p$ independent from $\epsilon$, the distribution of $X$ is absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^p$, the support of $f_X$ is convex, $\text{Var}(X) = \Sigma_X$ is positive definite, $\mathbb{E}(\epsilon) = 0$, $\text{Var}(\epsilon) = \eta^2 < \infty$.

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

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

(A.4). $\text{supp}(f_X)$ is compact.

The intersection of all subspaces for which (2) holds is called the mean subspace ([26, Def. 8.1]) and the intersection of all subspaces for which $F(Y \mid X) = F(Y \mid B^T X)$ holds is called the central subspace ([26, Def. 2.1]). The mean and central subspaces agree in this model since $s_0$ where $r_1 = V^T(x - s_0) \in \mathbb{R}^q, r_2 = U^T(x - s_0) \in \mathbb{R}^{p-q}$.

In the sequel, we refer to the following assumptions as needed. Lengthy proofs are given in the Appendix.
is a proper conditional density of $X$ that is concentrated on $\{s_0 + \text{span}\{V\}\} \cap \text{supp}(f_X)$ by the concept of regular conditional probability [25] under assumption (A.2). The detailed justification is given in the Appendix, where we also show that under assumptions (A.1), (A.2) and (A.4), $\hat{L}(V, s_0)$ in (6) and $L(V)$ in (7) are well defined and continuous. Moreover,

$$\hat{L}(V, s_0) = \mu_2(V, s_0) - \mu_1(V, s_0)^2 + \eta^2$$

(9)

where

$$\mu_1(V, s_0) = \int_{\mathbb{R}^q} g(B^T s_0 + B^T V r_1)^l f_X(s_0 + V r_1) dr_1 = \frac{\ell^{(l)}(V, s_0)}{\ell^{(0)}(V, s_0)}$$

(10)

with

$$\ell^{(l)}(V, s_0) = \int_{\mathbb{R}^q} g(B^T s_0 + B^T V r_1)^l f_X(s_0 + V r_1) dr_1.$$ 

(11)

Theorem 1 provides the statistical motivation for the objective function (7) of the conditional vari-ance estimator for the span of $B$ in (1).

**Theorem 1.** Suppose $V = (v_1, \ldots, v_q) \in S(p, q)$ and $q \in \{1, \ldots, p\}$. Under assumptions (A.1), (A.2) and (A.4),

(a) for all $s_0 \in \text{supp}(f_X)$ and $V$ such that there exist $u \in \{1, \ldots, q\}$ with $v_u \in \text{span}\{B\}$, $\hat{L}(V, s_0) > \text{Var}(\epsilon) = \eta^2$ and $L(V) > \eta^2$.

(b) for all $s_0 \in \text{supp}(f_X)$ and span$\{V\} \perp \text{span}\{B\}$, $\hat{L}(V, s_0) = \eta^2$ and $L(V) = \eta^2$.

**Proof.** Let $s_0 \in \text{supp}(f_X)$ and $V = (v_1, \ldots, v_q) \in \mathbb{R}^{p \times q}$ so that $v_u \in \text{span}\{B\}$ for some $u \in \{1, \ldots, q\}$. To obtain (a), observe $X \in s_0 + \text{span}\{V\}$ implies $X = s_0 + P_V(X - s_0)$ and using (6) yields

$$\hat{L}(V, s_0) = \text{Var}\left(g(B^T X) \mid X \in s_0 + \text{span}\{V\}\right) + \text{Var}(\epsilon)$$

$$= \text{Var}\left(g(B^T s_0 + B^T P_V(X - s_0)) \mid X \in s_0 + \text{span}\{V\}\right) + \eta^2 > \eta^2$$

(12)

since $B^T P_V(X - s_0) \neq 0$ with probability 1, and therefore the variance term in (12) is positive. For $V$ such that $V$ and $B$ are orthogonal, $B^T P_V(X - s_0) = 0$ and (b) follows. Since $s_0$ is arbitrary yet constant, the statements for $L(V)$ follow.  

Theorem 1 has an intuitive geometrical interpretation for the proposed method. If $X$ is not random, the deterministic function $Y = g(B^T X)$ is constant in all directions orthogonal to $B$ and varies in all other directions. If randomness is introduced, as in model (1), then the variation in $Y$ stems only from $\epsilon$ in all directions orthogonal to $B$. In all other directions the variation comprises of the sum of the variation of $\epsilon$ and of $g(B^T X)$. In consequence, the objective function (7) captures the variation of $Y$ as $X$ varies in the column space of $V$ and is minimized in the directions orthogonal to $B$. For illustration, we return to the motivating example at the beginning of Section 2 in Figure 1. The direction of $B$ is the $x$-axis, whereas its orthogonal complement, $\{0, 1\}$, coincides with the $y$-axis, along which the response exhibits low variation arising solely from the error term $\epsilon$.

If we consider a shift point $s_0$, which is indicated by $\times$ in the plot, the arrow starting at $s_0$ in the left panel depicts the subspace $s_0 + \text{span}\{V\}$ for the direction $V \in \mathbb{R}^2$, with $V = B$, and the black dotted lines represent the slice $\{x \in \mathbb{R}^p : \|x - P_{s_0 + \text{span}\{V\}}x\| \leq b_n\}$. In this direction the variability
in the points as captured by (6) is maximal. On the other hand, in the right panel, the arrow depicts
the subspace \(s_0 + \text{span}\{V\}\) for the direction \(V = B^\perp\). Here, (6) is not only small but, importantly,
bounded below by the error variance, an unknown but fixed amount. Moreover, the shifting point is
not material. What matters only is whether \(V\) aligns with or is orthogonal to \(B\). Because of this, we
remove the shifting point by averaging the variability over all shifting points, which corresponds to the
objective function (7). In order to locate \(B\), we simply search for the direction \(V\) that minimizes (7).

Let

\[ V_q = \arg\min_{V \in S(p, q)} L(V). \]  

(13)

\(V_q\) is well defined as the minimizer of a continuous function over the compact set \(S(p, q)\). Neverthe-
less, \(V_q\) is not unique since, for all orthogonal \(O \in \mathbb{R}^{q \times q}\) such that \(OO^T = I_q\), \(L(VO) = L(V)\) as
\(L(V)\) depends on \(V\) only through \(\text{span}\{V\}\). It is though unique as an optimization over the Grass-
mann manifold (4) by the uniqueness of \(\text{span}\{B\}\) from assumption (A.1) and Proposition 6.4 in [10]
[see the remark after the proof of Theorem 3 in the Appendix].

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

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

(a) \(\text{span}\{V_q\} = \text{span}\{B\}^\perp\)
(b) \(\text{span}\{V_q\}^\perp = \text{span}\{B\}\)

Let \(B_{p-q}\) denote a basis of \(\text{span}\{V_q\}^\perp\), so that

\[ \text{span}\{B_{p-q}\} = \text{span}\{V_q\}^\perp. \]  

(14)

where \(V_q\) is given in (13). Equation (14) is the population estimation equation for the sufficient reduc-
tion space, \(\text{span}\{B\}\), in (1), which is motivated by Theorem 1 and Corollary 2 (b). When \(q = p - k\),
where \(k = \text{rank}(B)\) in (1), then the **conditional variance estimator** (CVE) obtains the population
\(\text{span}\{B\}\).

Alternatively, we can also target \(B\) directly by maximizing the objective function \(L(V)\). The down-
side of this approach is that \(X\) either needs to be standardized, or the conditioning argument needs to
be changed to \(X = s_0 + P_{\text{\scriptsize M}(\text{span}\{V\})}(X - s_0)\), where \(P_{\text{\scriptsize M}(\text{span}\{V\})}\) is the orthogonal projection
operator with respect to the inner product \((x, y)_M = x^T My\). In either case, the inversion of \(\Sigma_X\) is
required. Our choice of targeting the orthogonal complement avoids the inversion of \(\Sigma_X\), and the es-
timation algorithm in Section 5 can be applied to regressions with \(p > n\) or \(p \approx n\), where \(n\) denotes
the sample size. Additionally, targeting the complement has computational advantages. The dimension
of the search space \(\text{span}\{V_q\}^\perp\) is \(p - q\), smaller than the dimension of the direct target space in (14)
when \(q = p - k\) for small \(k\), which is the appropriate setting in a dimension reduction context.

3. Estimation

We defined the estimation equation at the population level in (14). To calculate the conditional variance
estimator (CVE) from a sample from model (1), we replace the objective function, (7), in (13) by an
estimate. We start by describing the estimation of the interim objective function (6) and the final target
function (7) in this section. The definition of CVE is given in Section 3.2. A nonparametric estimation
method will be used to estimate the function in (6). The replacement of the unknown \(g\) and \(f_X\) in (10) by standard nonparametric kernel estimates is unsuitable for the sufficient dimension reduction task since the goal is to avoid the curse of dimensionality of nonparametric estimation over a high input dimension. Therefore, we opted to use a kernel estimation approach that takes into account the structure of the conditioning subspaces by using slicing and thus reducing the dimension over which the nonparametric smoothing takes place to \(p - q = k\), which is substantially smaller than \(p\) if \(k\) is small. After an estimate for (6) is obtained the final objective function (7) is estimated via its sample version.

Assume \((Y_i, X_i^T)_{i=1,...,n}\) is a random sample from model (1). For \(V \in S(p,q)\) and \(s_0 \in \text{supp}(f_X)\), we define

\[
d_i(V, s_0) = \|X_i - P_{s_0 + \text{span}\{V\}} X_i\|^2 = \|X_i - s_0\|^2 - \langle X_i - s_0, VV^T(X_i - s_0) \rangle
\]

\[
= \|(I_p - VV^T)(X_i - s_0)\|^2 = \|P_U(X_i - s_0)\|^2
\]

(15)

where \(\langle \cdot, \cdot \rangle\) is the usual inner product in \(\mathbb{R}^p\), \(P_V = VV^T\) and \(P_U = I_p - P_V\) using the orthogonal decomposition given by (5).

Let \(h_n \in \mathbb{R}_+\) be a sequence of bandwidths. We call the set

\[
S_{s_0, V} = \{x \in \mathbb{R}^p: \|x - P_{s_0 + \text{span}\{V\}} x\|^2 \leq h_n\}
\]

a slice that depends on both the shifting point \(s_0\) and the matrix \(V\), so that \(h_n\) represents the squared width of a slice around the subspace \(s_0 + \text{span}\{V\}\). It satisfies the following assumptions.

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

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

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

Let \(K\) be a function satisfying the following assumptions.

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

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

1. \(K(u) = 0\) for \(|u| > L_2\) and for all \(u, \tilde{u}\) it holds \(|K(u) - K(\tilde{u})| \leq L_1|u - \tilde{u}|\)
2. \(K(u)\) is differentiable with \(|\partial_u K(u)| \leq L_1\) and for some \(\nu > 1\) it holds \(|\partial_u K(u)| \leq L_1|u|^{-\nu}\) for \(|u| > L_2\)

Examples of functions that satisfy (K.1) and (K.2) include the Gaussian, \(K(z) = c\exp(-z^2/2)\), the exponential, \(K(z) = c\exp(-z)\), and the squared Epanechnikov kernel, \(K(z) = c\max\{(1 - z^2),0\}^2\) (i.e. polynomial kernels), where \(c\) is a constant. The rectangular, \(K(z) = cI(z \leq 1)\), does not fulfill (K.1) or (K.2) but is used for intuitive explanations. A list of kernel functions is given in [32, Table 1].
3.1. The estimator of $L(V)$ and its uniform convergence

**Definition 3.** For $i = 1, \ldots, n$, we define
\[
  w_i(V, s_0) = \frac{K \left( \frac{d_i(V, s_0)}{h_n} \right)}{\sum_{j=1}^{n} K \left( \frac{d_j(V, s_0)}{h_n} \right)}
\]
(16)

**Definition 4.** The sample based estimate of $\tilde{L}(V, s_0)$ is defined as
\[
  \tilde{L}_n(V, s_0) = \sum_{i=1}^{n} w_i(V, s_0) (Y_i - \bar{y}_1(V, s_0))^2 = \bar{y}_2(V, s_0) - \bar{y}_1(V, s_0)^2
\]
(17)

where $\bar{y}_l(V, s_0) = \sum_{i=1}^{n} w_i(V, s_0) Y_i^l$, $l = 1, 2$.

**Definition 5.** The estimate of the objective function $L(V)$ in (7) is defined as
\[
  L_n(V) = \frac{1}{n} \sum_{i=1}^{n} \tilde{L}_n(V, X_i),
\]
(18)

where each data point $X_i$ is a shifting point.

To obtain insight as to the choice of $\tilde{L}_n(V, s_0)$ in (17), let us consider the rectangular kernel, $K(z) = 1_{\{z \leq 1\}}$. In this case, $\tilde{L}_n(V, s_0)$ computes the empirical variance of the $Y_i$'s corresponding to the $X_i$'s that are no further than $\sqrt{h_n}$ away from the affine space $s_0 + \text{span}\{V\}$; i.e.,
\[
  d_i(V, s_0) = ||X_i - P_{s_0 + \text{span}\{V\}} X_i||^2 \leq h_n.
\]
If a smooth kernel is used, such as the Gaussian in our simulation studies, then $\tilde{L}_n(V, s_0)$ is also smooth, which allows the computation of gradients required to solve the optimization problem.

In Theorem 3 we state the conditions under which $L_n(V)$ in (18) converges uniformly to its population counterpart in (7). This result leads to the consistency of our estimator.

**Theorem 3.** Let $\tilde{a}_n^2 = \log(n)/n$. Under assumptions (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1), $\tilde{a}_n^2 = \log(n)/n h_n^{(p-q)/2} = o(1)$, and $a_n/h_n^{(p-q)/2} = O(1)$,
\[
  \sup_{V \in \text{span}\{B\}} |L_n(V) - L(V)| \rightarrow 0 \text{ in probability as } n \rightarrow \infty
\]
(19)

3.2. The Conditional Variance Estimator

We now define the estimator we propose for the mean subspace $\text{span}\{B\}$ in (1). Our main theoretical result in Theorem 4 establishes the consistency of our estimator.

**Definition 6.** The Conditional Variance Estimator $\tilde{B}_{p-q}$ is any basis of $\text{span}\{\tilde{V}_q\}^\perp$, where $\tilde{V}_q = \arg\min_{V \in \text{span}\{\tilde{V}_q\}} L_n(V)$. 
Theorem 4. Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1), \(a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)\), and \(a_n/h_n^{(p-q)/2} = O(1)\), \(\text{span}\{B_k\}\), where \(k = \text{rank}(B)\), is a consistent estimator for \(\text{span}\{B\}\) in model (1). That is, \(\|P_{B_h} - P_B\| \to 0\) in probability as \(n \to \infty\).

3.3. Weighted estimation of \(L(V)\)

The set of points \(\{x \in \mathbb{R}^p : \|x - P_{s_0 + \text{span}(V)}x\|^2 \leq h_n\}\) represents a slice in the subspace of \(\mathbb{R}^p\) about \(s_0 + \text{span}\{V\}\). The estimation of \(L(V)\) involves two types of weights:

(a) Within a slice. The weights are defined in (16) and are used to calculate (17).
(b) Between slices. Equal weights \(1/n\) are used to calculate (18).

The choice of weights can be potentially influential. In this section, we adapt the estimation of the target function (7) to account for the usual lack of balance in number of data points among different slices by changing the between slice weights. This can be realized by altering (18) to

\[
L_n^{(w)}(V) = \sum_{i=1}^{n} \tilde{w}(V, X_i) \tilde{L}_n(V, X_i), \quad \text{with} \quad (20)
\]

where

\[
\tilde{w}(V, X_i) = \frac{\sum_{j=1}^{n} K(d_j(V, X_i)/h_n) - 1}{\sum_{l,u=1}^{n} K(d_l(V, X_u)/h_n) - n} = \frac{\sum_{j=1,j\neq i}^{n} K(d_j(V, X_i)/h_n)}{\sum_{l,u=1,l\neq u}^{n} K(d_l(V, X_u)/h_n)} \quad (21)
\]

For example, if a rectangular kernel is used, \(\sum_{j=1,j\neq i}^{n} K(d_j(V, X_i)/h_n)\) is the number of \(X_j\) (\(j \neq i\)) points in the slice corresponding to \(\tilde{L}_n(V, X_i)\). The larger the number of \(X_j\) points in a slice, the higher weight it gets. That is, the more observations we use for estimating \(L(V, X_i)\) the better its accuracy. The denominator in (21) guarantees the weights \(\tilde{w}(V, X_i)\) sum up to one.

If (18) is replaced by (20) in Definition 6, the resulting estimator is called weighted conditional variance estimator.

3.4. Bandwidth selection

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

Lemma 5. Let \(M\) be a \(p \times p\) positive definite matrix. Then, \(tr(M)/p = \arg\min_{s > 0} \|M - sI_p\|\)\( (22)\)

Proof. Let \(U\) be the \(p \times p\) matrix whose columns are the eigenvectors of \(M\) corresponding to its eigenvalues \(\lambda_1 \geq \ldots \geq \lambda_p > 0\). Then, \(M = U \text{diag}(\lambda_1, \ldots, \lambda_p) U^T\), which implies \(\|M - sI_p\|^2 = \|\text{diag}(\lambda_1 - s, \ldots, \lambda_p) - sI_p\|^2 = \sum_{l=1}^{p} (\lambda_l - s)^2\). Taking the derivative with respect to \(s\), setting it to 0 and solving for \(s\) obtains (22), since \(\sum_{l=1}^{p} \lambda_l = tr(M)\). \(\square\)
If the predictors are multivariate normal, their joint density is approximated by \( N(\mu_X, \sigma^2 I_p) \) by Lemma 5, with \( \sigma^2 = \text{tr}(\Sigma_X) / p \). This results in no bandwidth dependence on \( \sigma \) and leads to a rule for bandwidth selection, as follows.

Under \( X \sim N_p(\mu_X, \sigma^2 I_p) \), \( \bar{X}_i = X_i - X_j \sim N_p(0, 2\sigma^2 I_p) \) for \( i \neq j \), where we suppress the dependence on \( j \) for notational convenience. Since all data are used as shifting points, \( d_i(V, X_j) = \|X_i - X_j\|^2 - (X_i - X_j)^T V V^T (X_i - X_j) = \|\bar{X}_i\|^2 - \bar{X}_i^T V V^T \bar{X}_i \). Let

\[
\text{nObs} = E\left( \# \{i \in \{1, \ldots, n\} : \bar{X}_i \in \text{span}_h \{V\} \} \right) = 1 + (n - 1)\mathbb{P}(d_1(V, X_2) \leq h) = 1 + (n - 1)\mathbb{P}(\|\bar{X}\|^2 - \bar{X}^T V V^T \bar{X} \leq h) \tag{23}
\]

where \( \text{span}_h \{V\} = \{x \in \mathbb{R}^p : \|x - P_{\text{span}(V)} x\|^2 \leq h \} \) and \( \bar{X} = X - X^* \), with \( X^* \) an independent copy of \( X \). nObs is the expected number of points in a slice. Given a user specified value for nObs, \( h \) is the solution to (23).

Let \( x \in \mathbb{R}^p \). For any \( V \in S(p, q) \) in (3), there exists an orthonormal basis \( U \in \mathbb{R}^{p \times (p-q)} \) of \( \text{span}(V)^\perp \) such that \( x = VR_1 + UR_2 \), by (5). Then, \( \tilde{X} = VR_1 + UR_2 \), with \( R_1 = V^T \tilde{X} \sim N(0, 2\sigma^2 I_{p-q}) \), \( R_2 = U^T \tilde{X} \sim N(0, 2\sigma^2 I_{p-q}) \), and \( \tilde{X}^T V V^T \tilde{X} = \|R_1\|^2 \) and \( \|\tilde{X}\|^2 = \|R_1\|^2 + \|R_2\|^2 \). Therefore,

\[
\mathbb{P}\left( \|\tilde{X}\|^2 - \tilde{X}^T V V^T \tilde{X} \leq h \right) = \mathbb{P}(\|R_2\|^2 \leq h) = \chi_{p-q} \left( \frac{h}{2\sigma^2} \right), \tag{24}
\]

where \( \chi_{p-q} \) is the cumulative distribution function of a chi-squared random variable with \( p - q \) degrees of freedom. Plugging (24) in (23) obtains

\[
n\text{Obs} = 1 + (n - 1)\chi_{p-q} \left( \frac{h}{2\sigma^2} \right). \tag{25}
\]

Solving (25) for \( h \) and Lemma 5 yield

\[
h_{\text{nObs}} = \chi_{p-q}^{-1} \left( \frac{n\text{Obs} - 1}{n - 1} \right) \frac{2\text{tr}(\Sigma_X)}{p}, \tag{26}
\]

where \( \Sigma_X = \sum_i(X_i - \bar{X})(X_i - \bar{X})^T / n \) and \( \bar{X} = \sum_i X_i / n \).

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

Alternatively, a plug-in bandwidth based on rule-of-thumb rules of the form \( c s n^{-1/(4+k)} \), where \( s \) is an estimate of scale and \( c \) a number close to 1, such as Silverman’s \( c = 1.06 \), \( s = \text{standard deviation} \) or Scott’s \( c = 1.0, s = \text{standard deviation} \), used in nonparametric density estimation [see [33]], is

\[
h_n = 1.2 \left[ \frac{2\text{tr}(\Sigma_X)}{p}(n^{-1/(4+p-q)}) \right]^2. \tag{27}
\]

The term \( 2\text{tr}(\Sigma_X)/p \) can be interpreted as the variance of \( X_i - X_j \) and \( p - q \) is the true dimension \( k \). We use \( c = 1.2 \) based on empirical evidence from simulations. Since both (26) and (27) yield satisfactory results, we opted against cross validation for bandwidth selection because of the computational burden involved, and used the bandwidth in (27) in simulations and data analyses.
4. A study of the behaviour of \( L_n(V) \)

In this section we demonstrate how the sample version (18) of the objective function (7) estimates the orthogonal complement of \( B \) in (1) via an example. We refer again to Figure 1 and the toy example \( Y = X_1 + \epsilon, X = (X_1, X_2)^T \sim N(0, I_2) \), with \( \epsilon \sim N(0, \eta^2) \) independent of \( X \), we used in Section 2 to provide an intuitive motivation for the proposed approach. Here we use it to describe the estimation algorithm visually.

For each point \( X_i, d_i(V, s_0) \) in (15) is its squared distance from the subspace \( s_0 + \text{span}\{V\} \). If the rectangular kernel \( K(z) = 1_{\{|z| \leq 1\}} \) is used, \( K(d_i(V, s_0)/h_n) \) is 0 for points outside the slice and 1 inside. Then, \( \hat{L}_n(V, s_0) \) in (17) calculates the empirical variance of the \( Y_i \) values whose corresponding \( X_i \) values fall into the slice. In the left panel of the graph, \( V \) is aligned with \( B \), which yields \( \hat{L}_n(V, s_0) = 1.034 \), and the color in the slice is very heterogeneous with high variation. In contrast, in the right panel \( V \) is orthogonal to \( B \), resulting in \( \hat{L}_n(V, s_0) = 0.11 \), and the color in the slice is very homogeneous with low variation.

Since one slice uses only a fraction of the data and the variation in one slice only depends on the alignment of the direction \( V \) to \( B \) but not on the shifting point \( s_0 \), as long as the slice is not too sparsely populated by samples, it is useful to average (17) over different shifting points \( s_0 \) in order to use all data available. A convenient choice for shift points are the datapoints \( X_i \), therefore we average over all \( X_i \) in (18) to form the final estimate of the objective function.

The population quantity in (6) results from making the width of the slice, \( h_n \), infinitesimally small and \( X \) ranging only in the subspace \( s_0 + \text{span}\{V\} \), so that infinitely many \( X_i \)'s lie exactly on the line spanned by the black arrow. In this case, in the right panel, (6) calculates \( \text{Var}(\epsilon) = \eta^2 \) and in the left panel \( \text{Var}(Y | X \in s_0 + \text{span}\{V\}) + \text{Var}(\epsilon) \). This demonstrates Theorem 1.

In this toy model we can calculate (7) explicitly via (8) and (10). We do so for an arbitrary \( \Sigma_X \) to demonstrate that the minimizer of (7) is always orthogonal to \( B \) but the maximizer can be influenced by the covariance matrix. Under the specifications of this toy model, (10) becomes

\[
\mu(V, s_0) = \int_\mathbb{R} (B^T s_0 + B^T V r)^T \psi_{X|\in s_0+\text{span}\{V\}}(r) dr
\]

(28)

Dropping the terms that do not contain \( r \) in (8) yields

\[
\hat{f}_{X|\in s_0+\text{span}\{V\}}(r) \propto f_{X}(s_0 + V r) \propto \exp \left( -\frac{1}{2} (s_0 + r V)^T \Sigma_X^{-1} (s_0 + r V) \right)
\]

\[
\propto \exp \left( -\frac{1}{2} (2r V^T \Sigma_X^{-1} s_0 + r^2 V^T \Sigma_X^{-1} V) \right) = \exp \left( -\frac{1}{2\sigma^2} (2r \sigma^2 V^T \Sigma_X^{-1} s_0 + r^2) \right)
\]

\[
\propto \exp \left( -\frac{1}{2\sigma^2} (r - \alpha)^2 \right),
\]

(29)

where \( \sigma^2 = 1/(V^T \Sigma_X^{-1} V) \), \( \alpha = -\sigma^2 V^T \Sigma_X^{-1} s_0 \) and the symbol \( \propto \) stands for proportional to. Letting \( \psi(z) \) denote the density of a standard normal variable, (29) obtains

\[
\hat{f}_{X|\in s_0+\text{span}\{V\}}(r) = \frac{1}{\sigma} \psi \left( \frac{r - \alpha}{\sigma} \right)
\]

(30)

for \( V, s_0 \in \mathbb{R}^{2 \times 1} \). Inserting (30) in (28) yields

\[
\int_\mathbb{R} (B^T s_0 + B^T V r)^T \frac{1}{\sigma} \psi \left( \frac{r - \alpha}{\sigma} \right) dr = \begin{cases} 
B^T s_0 + B^T V \alpha, & l = 1 \\
(B^T s_0)^2 + 2(B^T s_0)(B^T V) \alpha + (B^T V)^2 (\sigma^2 + \alpha^2), & l = 2
\end{cases}
\]
Using (9), (6) and (7), yields
\[ \tilde{L}(V, s_0) = \mu_2(V, s_0) - \mu_1(V, s_0)^2 + \eta^2 = (B^T V)^2 \sigma^2 + \eta^2, \]
so that
\[ L(V) = E(\tilde{L}(V, X)) = (B^T V)^2 \sigma^2 + \eta^2 = \frac{(B^T V)^2}{V^T \Sigma^{-1} V} \]
(31)

From (31) we can easily see that \( L(V) \) attains its minimum at \( V \perp B \). Also, if \( \Sigma_X = I_2 \), the maximum of \( L(V) \) is attained at \( V = B \).

The true \( L(V(\theta)) \) and its estimates \( L_n(V(\theta)) \) are plotted for samples of different sizes \( n \) in Figure 2. \( L_n(V(\theta)) \) approximates \( L(V) \) fast and attains its minimum at the same value as \( L(V) \) even for \( n = 10 \).

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

5. Optimization Algorithm

A Stiefel manifold optimization algorithm is used to obtain the solution of the sample version of the optimization problem (13). To calculate \( \hat{V}_{q \times p} \) in (6), a curvilinear search is carried out [35, 1], which is similar to gradient descent. First an arbitrary starting value \( V(0) \) is selected by drawing a \( p \times q \) matrix from the invariant measure; i.e., the distribution that corresponds to the uniform, on \( S(p, q) \), see [9]. The \( Q \)-component of the QR decomposition of a \( p \times q \) matrix with independent standard normal entries follows the invariant measure [8]. The step-size \( \tau > 0 \), the step size reduction factor \( \gamma \in (0, 1) \), and tolerance \( \text{tol} > 0 \) are fixed at the outset.
Result: $V^{(\text{end})}$

Initialize: $V^{(0)}$, $\tau = 1$, $\text{tol} = 10^{-3}$, $\gamma = 0.5$ error = tol + 1, maxit = 50, count = 0;

while error > tol and count $\leq$ maxit do
  • $G = \nabla V \hat{L}_n(V^{(j)}) \in \mathbb{R}^{p \times q}$, $W = GV^T - VG^T$
  • $V^{(j+1)} = (I_p + \tau W)^{-1}(I_p - \tau W)V^{(j)}$
  • error = $\|V^{(j)}V^{(j)^T} - V^{(j+1)}V^{(j+1)^T}\|/\sqrt{2q}$

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

end

Algorithm 1: Curvilinear search

Under mild regularity conditions on the objective function, [35] showed that the sequence generated by the algorithm converges to a stationary point if the Armijo-Wolfe conditions [31] are used for determining the stepsize $\tau$.

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

The algorithm is repeated for $m$ arbitrary $V^{(0)}$ starting values drawn from the invariant measure on $S(p, q)$. Among those, the value at which $L_n$ in (18) is minimal is selected as $\hat{V}_q$.

The algorithm requires the computation of the gradient of $L_n(V)$ in (18) or (20). We compute the gradient of the objective function for the Gaussian kernel in Theorems 6 and 7. The Gaussian kernel is the default kernel we use in the implementation of the estimation algorithm in the R package CVarE.

Theorem 6. Let $K(z) = \exp(-z^2/2)$ be the Gaussian kernel. Then, the gradient of $\hat{L}_n(V, s_0)$ in (17) is given by

$$\nabla V \hat{L}_n(V, s_0) = \frac{1}{\hat{n}} \sum_{i=1}^{n} (\hat{L}_n(V, s_0) - (Y_i - \hat{y}_1(V, s_0))^2) w_i d_i \nabla V d_i(V, s_0) \in \mathbb{R}^{p \times q},$$

and the gradient of $L_n(V)$ in (18) is

$$\nabla V L_n(V) = \frac{1}{n} \sum_{i=1}^{n} \nabla V \hat{L}_n(V, X_i).$$

with $w_i = w(V, X_i)$ in (16).
The weighted version of conditional variance estimation in Section 3.3 is expected to increase the accuracy of the estimator for unevenly spaced data. When (20) and the gradient in (32) are used in the optimisation algorithm, we refer to the estimator as weighted conditional variance estimation. If (20) and the gradient $\sum_{i=1}^{n} \hat{w}(V, X_i) \nabla \tilde{L}_n(V, X_i)$ are used; i.e., the first summand in (32) is dropped, we refer to it as partially weighted conditional variance estimation. For both, we replace $G$ in algorithm 1 with the corresponding gradient derived in Theorem 7.

**Theorem 7.** Let $K(z) = \exp (-z^2/2)$ be the Gaussian kernel. Then, the gradient of $L_n^{(w)}(V)$ in (20) is given by

$$\nabla V L_n^{(w)}(V) = \sum_{i=1}^{n} \left( \nabla V \hat{w}(V, X_i) \tilde{L}_n(V, X_i) + \hat{w}(V, X_i) \nabla \tilde{L}_n(V, X_i) \right),$$

where $\nabla V \tilde{L}_n(V, X_i)$ is given in Theorem 6. Furthermore,

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

with $\hat{w}_i = \hat{w}(V, X_i)$ in (21), $K_{j,i} = K(d_j(V, X_i)/h_n)$, and $d_{j,i} = d_j(V, X_i)$ given in (15).

CVE has computational complexity of $O(m \text{ maxit } pqn^2)$, which can be seen in both Theorem 6 and Theorem 7, since, for each entry of the $p \times q$ dimensional gradient, a double sum with both indices ranging over 1, \ldots, $n$ is evaluated, and, for the $m$ starting values, at most maxit iterations are used. The formulas of the gradient are straightforward to implement, yet there is no closed form solution as $V$ is inside the nonlinear kernel.

### 6. Simulation studies

We compare the estimation accuracy of conditional variance estimation with the forward model based sufficient dimension reduction methods, mean outer product gradient estimation (meanOPG), mean minimum average variance estimation (meanMAVE) [19], refined outer product gradient (rOPG), refined minimum average variance estimation (rmave) [37, 26], and principal Hessian directions (pHd) [28, 14], and the inverse regression based methods, sliced inverse regression (SIR) [27] and sliced average variance estimation (SAVE) [16]. The dimension $k$ is assumed to be known throughout.

We report results for conditional variance estimation using the “plug-in” bandwidth in (27) and three different conditional variance estimation versions, CVE, wCVE, and rCVE. CVE is obtained by using $m = 10$ arbitrary starting values in the optimization algorithm and optimizing (18) as described in Section 5. rCVE, or refined weighted CVE, is obtained by setting the starting value $V^{(0)}$ at the optimizer of CVE, and using (20) in the optimization algorithm in Section 5 with the partially weighted gradient as described in Section 3.3. wCVE, or weighted CVE, is obtained by optimizing (20) with partially weighted gradient as described in Sections 3.3 and 5. Methods rOPG and rmave refer to the original refined outer product gradient and refined minimum average variance estimation algorithms published in [37]. They are implemented using the R code in [26] with number of iterations $\text{nit} = 25$, since the algorithm is seen to converge by 25. The dr package is used for the SIR, SAVE and pHd calculations, and the MAVE package for mean outer product gradient estimation (meanOPG) and mean
minimum average variance estimation (meanMAVE). CVE is implemented in the R package CVarE [23], available at https://cran.r-project.org.

Table 1 lists the seven models (M1-M7) we consider. Throughout, we set $p = 20$. For M1-M5, $b_1 = (1,1,1,1,1,0,...,0)^T/\sqrt{6}$, $b_2 = (1,-1,-1,-1,1,-1,0,...,0)^T/\sqrt{6} \in \mathbb{R}^p$. For M6, $b_1 = e_1$, $b_2 = e_2$ and $b_3 = e_6$, and for M7 $b_1, b_2, b_3$ are the same as in M6 and $b_4 = e_3$, where $e_j$ denotes the $p$-vector with $j$th element equal to 1 and all others are 0. The error term $\epsilon$ is independent of $X$ for all models. In M2, M3, M4, M5 and M6, $\epsilon \sim N(0,1)$. For M1 and M7, $\epsilon$ has a generalized normal distribution $GN(\alpha, b, c)$ with density $f_\epsilon(z) = c/2\sqrt{\pi}e^{-\epsilon/\alpha}$, see [30] with location $0$ and shape-parameter $0.5$ for M1, and shape-parameter $1$ for M7 (Laplace distribution). For both the scale-parameter is chosen such that $\text{Var}(\epsilon) = 0.25$.

| Name | Model | $X$ distribution | $\epsilon$ distribution | $k$ | $n$ |
|------|-------|-----------------|--------------------------|-----|-----|
| M1   | $Y = \cos(b_1^T X) + \epsilon$ | $X \sim N_p(0, \Sigma)$ | $GN(0, \sqrt{0.5})$ | 1  | 100 |
| M2   | $Y = \cos(b_2^T X) + 0.5\epsilon$ | $X \sim N_p(0, \Sigma)$ | $N(0,1)$ | 1  | 100 |
| M3   | $Y = 2\log(|b_3^T X| + 2) + 0.5\epsilon$ | $X \sim N_p(0, \Sigma)$ | $N(0,1)$ | 1  | 100 |
| M4   | $Y = (b_4^T X)/(0.5 + (1 + b_5^T X)^2) + 0.5\epsilon$ | $X \sim N_p(0, \Sigma)$ | $N(0,1)$ | 2  | 200 |
| M5   | $Y = \cos(\pi b_6^T X)/(b_7^T X + 1)^2 + 0.5\epsilon$ | $X \sim U([0,1]^p)$ | $N(0,1)$ | 2  | 200 |
| M6   | $Y = (b_7^T X)^2 + (b_8^T X)^2 + (b_9^T X)^2 + 0.5\epsilon$ | $X \sim N_p(0, \Sigma)$ | $N(0,1)$ | 3  | 200 |
| M7   | $Y = (b_7^T X)/(b_2^T X)^2 + (b_8^T X)/(b_4^T X) + \epsilon$ | $X \sim N_p(0, \Sigma)$ | $GN(0, \sqrt{1/(0.5)})$ | 4  | 400 |

The variance-covariance structure of $X$ in models M1 and M4 satisfies $\Sigma_{i,j} = 0.5|j-i|$ for $i, j = 1, \ldots, p$. In M5, $X$ is uniform with independent entries on the $p$-dimensional hyper-cube. In M7, $X$ is multivariate $t$-distributed with 3 degrees of freedom. The link functions of M4 and M7 are studied in [37], 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, $Z \sim 2 \text{Bernoulli}(p_{mix}) - 1 \in \{-1,1\}$, where $1_q = (1,1,\ldots,1)^T \in \mathbb{R}^q$, mixing probability $p_{mix} \in [0,1]$ and dispersion parameter $\lambda > 0$. For $0 < p_{mix} < 1$, $X$ has a mixture normal distribution, where $p_{mix}$ is the relative mode height and $\lambda$ is a measure of mode distance.

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

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

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

In Figure 3, box-plots for all combinations of $p_{mix} \in \{0.3,0.4,0.5\}$ and $\lambda \in \{0,0.5,1,1.5\}$ are presented. The reference methods are restricted to meanOPG and meanMAVE, since the others are not
competitive. Conditional variance estimation performs better than all competing methods and is the only method with consistently smaller errors when the two modes are further apart ($\lambda \geq 1$) regardless of the mixing probability $p_{mix}$. 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, meanMAVE’s performance deteriorates as the bimodality of the predictor distribution becomes more distinct. In contrast, conditional variance estimation is unaffected. and appears to have an advantage over meanMAVE when the predictors have mixture distributions, the link function is even about the midpoint of the two modes, and $\mathbf{B}$ is not orthogonal to the line connecting the two modes. Conditional variance estimation is the only method that estimates the mean subspace reliably in model M2 ($\text{err} \approx 0.4$ to 0.5), whereas meanMAVE misses it completely ($\text{err} \approx 1$). These results indicate that conditional variance estimation is often approximately on par, and can perform much better than meanMAVE depending on the predictor distribution and the link function.

Table 2. Mean and standard deviation of estimation errors

| Model | CVE | wCVE | rCVE | meanOPG | rOPG | meanMAVE | rmave | pHd | sir | save |
|-------|-----|------|------|---------|------|-----------|-------|-----|-----|------|
| M1    |     |      |      | 0.3827  | 0.4414 | 0.4051    | 0.6220 | 0.9876 | 0.8278 | 0.9875 | 0.9788 |
|       | sd  |      |      | 0.1269  | 0.1595 | 0.1329    | 0.1879 | 0.0223 | 0.1206 | 0.0243 | 0.0334 |
| M2    |     |      |      | 0.4572  | 0.4992 | 0.4658    | 0.8987 | 0.9332 | 0.9242 | 0.9000 | 0.9783 | 0.9781 |
|       | sd  |      |      | 0.1038  | 0.1524 | 0.0989    | 0.0908 | 0.0683 | 0.0983 | 0.0735 | 0.0278 | 0.0318 |
| M3    |     |      |      | 0.6282  | 0.7509 | 0.6371    | 0.7847 | 0.9644 | 0.9674 | 0.6964 | 0.9647 | 0.9519 |
|       | sd  |      |      | 0.2354  | 0.2262 | 0.2181    | 0.2201 | 0.0667 | 0.2435 | 0.0609 | 0.1626 | 0.0587 | 0.0650 |
| M4    |     |      |      | 0.5663  | 0.5897 | 0.5554    | 0.4071 | 0.4026 | 0.3905 | 0.3772 | 0.5824 | 0.9727 |
|       | sd  |      |      | 0.1239  | 0.1246 | 0.1298    | 0.0814 | 0.0609 | 0.0997 | 0.0584 | 0.0662 | 0.0951 | 0.0202 |
| M5    |     |      |      | 0.4429  | 0.5604 | 0.4779    | 0.4058 | 0.3737 | 0.3929 | 0.3750 | 0.7329 | 0.6374 | 0.9730 |
|       | sd  |      |      | 0.0891  | 0.1233 | 0.0976    | 0.1022 | 0.0680 | 0.0894 | 0.0871 | 0.0832 | 0.0968 | 0.0186 |
| M6    |     |      |      | 0.3828  | 0.3027 | 0.3230    | 0.1827 | 0.4632 | 0.1656 | 0.4863 | 0.4978 | 0.9129 | 0.8236 |
|       | sd  |      |      | 0.1006  | 0.0748 | 0.1098    | 0.0289 | 0.1717 | 0.0252 | 0.1676 | 0.0601 | 0.0420 | 0.0518 |
| M7    |     |      |      | 0.6856  | 0.5050 | 0.5651    | 0.5694 | NA     | 0.5824 | 0.9727 | 0.8536 | 0.8133 | 0.8699 |
|       | sd  |      |      | 0.0588  | 0.0862 | 0.0879    | 0.1122 | NA     | 0.1271 | 0.0354 | 0.0341 | 0.0342 |

Furthermore we estimate the dimension $k$ via cross-validation, following the approach in [37], with

$$\hat{k} = \arg\min_{l=1,\ldots,p} CV(l) = \arg\min_{l=1,\ldots,p} \frac{\sum_{i}(Y_i - \hat{g}^{-i}(\hat{\mathbf{B}}_l^T \mathbf{X}_i))^2}{n},$$

(33)

where $\hat{g}^{-i}(\cdot)$ is computed from the data $(Y_j, \hat{\mathbf{B}}_l^T \mathbf{X}_j)_{j=1,\ldots,n;j\neq i}$ using multivariate adaptive regression splines [18] in the R-package mda, and $\hat{\mathbf{B}}_l = \hat{\mathbf{V}}_{p-l}$ is any basis of the orthogonal complement of $\hat{\mathbf{V}}_{p-l} = \arg\min_{\mathbf{V} \in S(p,p-l)} L_n(\mathbf{V})$. For a given $l$, we calculate $\hat{\mathbf{B}}_l$ from the whole data set and predict $Y_i$ by $\hat{Y}_{i,l} = \hat{g}^{-i}(\hat{\mathbf{B}}_l^T \mathbf{X}_i)$. For $l = p$, $\hat{\mathbf{B}}_p = \mathbf{I}_p$. The results for the seven models are reported in Table 3. The CVE based dimension estimation is the most accurate in models M1, M2, M3, and M6 and differs slightly from that of MAVE in M7. MAVE performs better in M4 and M5, completely misses the true
dimension in M2 and misses it most of the time in M3. Thus, the dimension estimation performance of CVE and MAVE agrees with the estimation accuracy of the true subspace in Table 2. CVE estimates the dimension more accurately even in model M6, where it exhibits worse subspace estimation performance, and overall appears to be more accurate.

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 |

We carried out many simulation experiments for an array of combinations of link functions, sufficient reduction matrices \( B \) and their ranks, as well as predictor and error distributions. All reported and unreported results indicate that the difference in performance of the two methods, CVE and mean MAVE, can be attributed to both the form of the link function and the marginal predictor distribution. Both CVE and MAVE minimise conditional variances with different conditioning arguments. MAVE minimises the objective function

\[
T(V) = \mathbb{E} \left( Y - \mathbb{E}(Y \mid V^T X) \right) = \mathbb{E} \left( \text{Var}(Y \mid V^T X) \right),
\]

hence it conditions on the projection \( V^T X \), whereas CVE conditions on \( X \) only ranging in an affine subspace \( s_0 + \text{span}\{V\} \) [see (7)]. A particular feature of CVE is that it does not require the estimation and inversion of the variance-covariance matrix of the predictors, which typically engenders inaccuracy in estimation. We also observed that when the link function had a bounded first order derivative, CVE...
often outperformed mean MAVE across predictor distributions. In the opposite case, MAVE performed mostly better. Moreover, when the predictors have a bimodal distribution with well separated modes and the link function is even, regardless of whether its derivative is bounded, CVE outperforms mean MAVE. In the other settings for the generated data, both methods were roughly on par.

7. Real Data Analyses

Three data sets are analyzed: the Hitters data in the R package ISLR, which was also analyzed by [37], the Boston Housing data in the R package mlbench, and the Concrete data from the MAVE package. The reference method is meanMAVE from the MAVE package in R and the CVE is calculated using $m = 50$ and maxit = 10 in the optimization algorithm 1 in Section 5. The estimation of the dimension is based on (33) in Section 6.

Following [37], we remove 7 outliers from the Hitters data set leading to a sample size of 256. The response is $Y = \log$(salary) and the 16 continuous predictors are the game statistics of players in the Major League Baseball league in the seasons 1986 and 1987. Further information can be found in https://www.rdocumentation.org/packages/ISLR/versions/1.2/topics/Hitters.

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

The Concrete data set contains 1030 instances on 9 continuous variables The response is concrete compressive strength. Concrete strength is very important in civil engineering and is a highly nonlinear function of age and ingredients. The description of the variables can be found in https://www.rdocumentation.org/packages/MAVE/versions/1.3.10/topics/Concrete.

For all three data sets we standardize both the predictors and the response by subtracting the mean and rescaling column-wise so that each variable has unit variance. The data sets are analyzed using 10 fold cross-validation to calculate an unbiased estimate of the prediction error [34] for our method, CVE, and its main competitor meanMAVE using the MAVE package. The dimension for each method is estimated with (33) on the trainings set and we then fit a forward regression model on the training set replacing the original with the reduced predictors using multivariate adaptive regression splines [18] using the R package mda and calculate the prediction error on the test set for both methods. The dimension estimates of CVE and MAVE mostly disagree.

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

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

| Method | Hitters | Housing | Concrete |
|--------|---------|---------|----------|
| CVE    | 0.216   | 0.260   | 0.361    |
|        | (0.101) | (0.331) | (0.206)  |
| MAVE   | 0.203   | 0.299   | 0.417    |
|        | (0.083) | (0.382) | (0.348)  |
7.1. Hitters Data Analysis as in [37]

Additionally, we reconstruct the analysis of the Hitters data in [37], which does not account for the out-of-sample prediction error as in Section 7 but uses the whole sample for estimation of $B$ and its rank. Only the dimension $k$ is estimated with leave-one-out cross validation.

Table 5 reports the average cross validation mean squared error $CV(k)$ in (33) using the whole data set over $k = 1, \ldots, 5$. Both conditional variance estimation and mean minimum average variance estimation estimate the dimension to be 2.

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

We plot the response against the estimated directions in Figure 4. Both exhibit the same pattern:

![Figure 4. $Y$ against $\hat{b}_1^T X$ and $\hat{b}_2^T X$](image)

the response appears to be linear in one direction and quadratic in the second. The difference is that the linear pattern is clearer in the second CVE direction and the quadratic pattern exhibits increasing variance in the first MAVE direction.

Based on the scatterplots in Figure 4, we fit the same models for both. For conditional variance estimation, the fitted regression is

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

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

$$\hat{Y} = 0.39051 + 1.32529(\hat{b}_1^T X) - 0.55328(\hat{b}_1^T X)^2 + 0.49546(\hat{b}_2^T X)$$
with $R^2 = 0.7859$. Both models (34) and (35) have about the same fit as measured by $R^2$. The in sample performance of the two methods is practically the same for the Hitters data.

8. Discussion

In this paper the novel conditional variance estimator (CVE) for the mean subspace is introduced. We present its geometrical and theoretical foundation, show its consistency and propose an estimation algorithm with assured convergence. CVE requires the forward model (1), $Y = g(B^T X) + \epsilon$, holds and weak assumptions on the response and the covariates.

Minimum average variance estimation (MAVE) [37] is the only other sufficient dimension reduction method based on the forward model (1). It estimates the sufficient dimension reduction targeting both the reduction and the link function $g$ in (1). CVE targets only the reduction and does not require estimation of the link function, which may explain why it has an advantage over MAVE in some regression settings. For example, CVE exhibits similar performance across different link functions (cos, exp, etc) for fixed $\lambda$, whereas the performance of MAVE is very uneven for model M2 in Section 6. CVE is more accurate than MAVE when the link function is even and the predictor distribution is bimodal throughout our simulation studies. Moreover, CVE does not require the inversion of the predictor covariance matrix and can be applied to regressions with $p \approx n$ or $p > n$.

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

References

[1] P. A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. Mathematics of Computation - Math. Comput., 78, 12 2008.
[2] K. P. Adragni and R. D. 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.
[3] T. Amemiya. Advanced Econometrics. Harvard university press, 1985.
[4] S. N. Bernstein. Theory of Probability. Moscow, 1927.
[5] W. M. Boothby. An Introduction to Differentiable Manifolds and Riemannian Geometry. Academic Press, 2002.
[6] E. Bura, S. Duarte, and L. Forzani. Sufficient reductions in regressions with exponential family inverse predictors. Journal of the American Statistical Association, 111(515):1313–1329, 2016.
[7] E. Bura and L. Forzani. Sufficient reductions in regressions with elliptically contoured inverse predictors. Journal of the American Statistical Association, 110(509):420–434, 2015.
[8] Y. Chikuse. Invariant measures on Stiefel manifolds with applications to multivariate analysis, volume Volume 24 of Lecture Notes–Monograph Series, pages 177–193. Institute of Mathematical Statistics, Hayward, CA, 1994.
[9] Y. Chikuse. Statistics on Special Manifolds. Springer-Verlag New York, New York, 2003.
[10] R. D. Cook. Regression Graphics: Ideas for studying regressions through graphics. Wiley, New York, 1998.
[11] R. D. Cook. Fisher lecture: Dimension reduction in regression. Statist. Sci., 22(1):1–26, 02 2007.
[12] R. D. Cook and L. Forzani. Principal fitted components for dimension reduction in regression. Statistical Science, 23(4):485–501, 2008.
[13] R. D. Cook and L. Forzani. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 3 2009.

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

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

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

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

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

[19] W. Hang and Y. Xia. *MAVE: Methods for Dimension Reduction*, 2019. R package version 1.3.10.

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

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

[22] I.M. James, Cambridge University Press, and N.J. Hitchin. *The Topology of Stiefel Manifolds*. Cambridge books online. Cambridge University Press, 1976.

[23] D. Kapla and L. Fertl. *CVarE: Conditional Variance Estimator for Sufficient Dimension Reduction*, 2021. R package version 1.1.

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

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

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

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

[28] K. C. 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.

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

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

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

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

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

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

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

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

[37] Y. Xia, H. Tong, W. K. Li, and L. X. Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.
Acknowledgements

The authors gratefully acknowledge the support of the Austrian Science Fund (FWF P 30690-N35) and thank Daniel Kapla for his programming assistance and Tijana Levajkovic and Michael Messer for their thorough reading and suggestions. Daniel Kapla co-authored the CVarEst package that implements the proposed method. We also thank two reviewers and an associate editor, whose suggestions helped us to improve the paper.

Appendix

Justification for (8): Theorem 3.1 of [25] and the fact that $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, where $\mathcal{B}(\mathbb{R}^p)$ denotes the Borel sets on $\mathbb{R}^p$, is a Polish space guarantee the existence of the regular conditional probability of $X \mid X \in s_0 + \text{span}\{V\}$ [see also [17]]. Further, the measure is concentrated on the affine subspace $s_0 + \text{span}\{V\} \subset \mathbb{R}^p$ and is given by (8) by Definition 8.38 and Theorem 8.39 of [24] and the orthogonal decomposition (5).

Proof of (9): Since $X$ and $\epsilon$ in (1) are assumed to be independent, $\text{Var}(Y \mid X \in s_0 + \text{span}\{V\}) = \text{Var}(g(B^T X) \mid X \in s_0 + \text{span}\{V\}) + \text{Var}(\epsilon)$. Using (8) and $\text{Var}(Y \mid Z) = E(Y^2 \mid Z) - E(Y \mid Z)^2$, we obtain (9).

We let $\hat{g}(V, s_0, r) = g(B^T s_0 + B^T V r) f_X(s_0 + V r)$. The parameter integral (11) is well defined and continuous if (1) $\hat{g}(V, s_0, \cdot)$ is integrable for all $V \in S(p, q), s_0 \in \text{supp}(f_X)$, (2) $\hat{g}(\cdot, \cdot, r)$ is continuous for all $r$, and (3) there exists an integrable dominating function of $\hat{g}$ that does not depend on $V$ and $s_0$ [see [21, p. 101]].

Furthermore $t^{(l)}(V, s_0) = \int_K \hat{g}(V, s_0, r) dr$ for some compact set $K$, since $\text{supp}(f_X)$ is compact due to (A.4). The function $\hat{g}(V, s_0, r)$ is continuous in all inputs by the continuity of $g$ and $f_X$ by (A.2), and therefore it attains a maximum. In consequence, all three conditions are satisfied so that $t^{(l)}(V, s_0)$ is well defined and continuous.

Next $\mu_t(V, s_0) = t^{(l)}(V, s_0)/t^{(0)}(V, s_0)$ is continuous since $t^{(0)}(V, s_0) > 0$ for all $s_0 \in \text{supp}(f_X)$ by the continuity of $f_X$ and $\Sigma_X > 0$. Then, $\tilde{L}(V, s_0)$ in (9) is continuous, which results in $L(V)$ also being well defined and continuous by virtue of it being a parameter integral following the same arguments as above.

Next we establish the consistency of the conditional variance estimator. The uniform convergence in probability of the sample objective function in (18) is a sufficient condition for obtaining the consistency of $\hat{V}_q = \arg\min_{V \in S(p, q)} L_n(V)$, as uniform convergence in probability of a random function implies convergence in probability of the minimizer of $L_n(V)$ to the minimizer of the limit function. Let

$$t_n^{(l)}(V, s_0) = \frac{1}{n h_n^{(p-q)/2}} \sum_{i=1}^n K\left(\frac{d_i(V, s_0)}{h_n}\right) V_i$$

be the sample version of (11) for $l = 0, 1, 2$. The summands of $\tilde{L}_n$ in (17) can be expressed as

$$\tilde{y}_l(V, s_0) = \frac{t_n^{(l)}(V, s_0)}{t_n^{(0)}(V, s_0)}.$$
Next we state auxiliary lemmas whose proofs are provided in the supplementary material.

**Lemma 8.** Assume (A.4) and (K.1) hold. Let $Z_n(V, s_0) = \left(\sum_i g(X_i) \mathbb{I} K(d_i(V, s_0)/h_n)\right)/(nh_n^{(p-q)/2})$ for a continuous function $g$. Then,

$$
\mathbb{E}(Z_n(V, s_0)) = \int_{\text{supp}(f_X) \cap \mathbb{R}^{p-q}} K(||r_2||^2) \int_{\text{supp}(f_X) \cap \mathbb{R}^{q}} \tilde{g}(r_1, h_n^{1/2} r_2) dr_1 dr_2
$$

where $\tilde{g}(r_1, r_2) = g(s_0 + Vr_1 + Ur_2)f_X(s_0 + Vr_1 + Ur_2)$, $x = s_0 + Vr_1 + Ur_2$ in (5).

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

$$
\frac{\bar{b}^{(2,y),2} - \delta}{n h_n^{(p-q)/2}} \leq \text{Var}(t_n^{(l)}(V, s_0)) \leq \frac{\bar{b}^{(2,y),2} + \delta}{n h_n^{(p-q)/2}}
$$

for $n > n^*$ and $t_n^{(l)}(V, s_0)$, $l = 0, 1, 2$, in (36).

In Lemma 10, we show that $d_l(V, s_0)$ in (15) is Lipschitz in its inputs under assumption (A.4).

**Lemma 10.** Under assumption (A.4) there exists a constant $0 < C_2 < \infty$ such that for all $\delta > 0$ and $V, V_j \in S(p, q)$ with $||P_V - P_{V_j}|| < \delta$ and for all $s_0, s_j \in \text{supp}(f_X) \subset \mathbb{R}^p$ with $||s_0 - s_j|| < \delta$,

$$
|d_l(V, s_0) - d_l(V_j, s_j)| \leq C_2 \delta
$$

for $d_l(V, s_0)$ given by (15).

The proofs of Theorems 3 and 11 require the Bernstein inequality [4]: Let $Z_1, Z_2, \ldots$ be an independent sequence of bounded random variables $|Z_i| \leq b$. Let $S_n = \sum_{i=1}^n Z_i$, $E_n = \mathbb{E}(S_n)$ and $V_n = \text{Var}(S_n)$. Then,

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

(38)

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

$$
|K(u) - K(u')| \leq K^*(u') \delta
$$

(39)

for all $u, u'$ with $|u - u'| < \delta \leq L_2$ and $K^*(\cdot)$ is a bounded and integrable kernel function [see 20]. Specifically, if condition (1) of (K.2) holds, then $K^*(u) = L_11_{|u| \leq 2L_2}$. If (2) holds, then $K^*(u) = L_11_{|u| \leq 2L_2} + 1_{|u| > 2L_2}|u - L_2|^{-\nu}$.

Let $A = S(p, q) \times \text{supp}(f_X)$ and by a slight abuse of notation, we generically denote constants by $C$. In Theorems 11 and 12 we show that the variance and bias terms of (36) vanish uniformly in probability, respectively.

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

$$
\sup_{V \times s_0 \in A} \left|f_n^{(l)}(V, s_0) - \mathbb{E}(f_n^{(l)}(V, s_0)) \right| = O_p(a_n) \quad \text{for} \quad l = 0, 1, 2
$$

(40)
**Remark.** If we assume $|Y| < M_2 < \infty$ almost surely, the requirement $a_n/h_n^{(p-q)/2} = O(1)$ for the bandwidth can be dropped and the truncation step of the proof of Theorem 11 can be skipped.

**Proof of Theorem 11.** The proof is organized in 3 steps: a truncation step, a discretization step by covering $A = S(p, q) \times \text{supp}(f_X)$, and application of Bernstein’s inequality (38).

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

$$t_n^{(l)}(V, s_0) = (1/nh_n^{(p-q)/2}) \sum_i K(||P_U(X_i - s_0)||^2/h_n)Y_i1\{|Y_i| \leq \tau_n\}$$

be the truncated version of (36) and $\bar{R}_n^{(l)} = (1/nh_n^{(p-q)/2})\sum_i |Y_i|^41\{|Y_i| > \tau_n\}$ be the remainder of (36). Therefore $R_n^{(l)}(V, s_0) = t_n^{(l)}(V, s_0) - t_n^{(l)}_{\text{trc}}(V, s_0) \leq M_1 \bar{R}_n^{(l)}$ due to (K.1) and

$$\sup_{V \times s_0 \in A} |t_n^{(l)}(V, s_0) - \mathbb{E}(t_n^{(l)}(V, s_0))| \leq M_1 (R_n^{(l)} + \mathbb{E}\bar{R}_n^{(l)})$$

$$+ \sup_{V \times s_0 \in A} |t_n^{(l)}_{\text{trc}}(V, s_0) - \mathbb{E}(t_n^{(l)}_{\text{trc}}(V, s_0))|$$

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

$$\mathbb{E}\bar{R}_n^{(l)} = \frac{1}{h_n^{(p-q)/2}} \mathbb{E}(|Y_i|^41\{|Y_i| > \tau_n\}) \leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|Y_i|^2)^2} \sqrt{\mathbb{P}(|Y_i|^4 > \tau_n)}$$

$$\leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|Y_i|^2)} \left(\frac{\mathbb{E}(|Y_i|^4)}{a_n^4}\right)^{1/2} = o(a_n)$$

(43)

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

Therefore the first two terms of the right hand side of (42) converge to 0 with rate $a_n$ by (43) and Markov’s inequality. From now to the end of the proof $Y_i$ will denote the truncated version $Y_i1\{|Y_i| \leq \tau_n\}$ and we do not distinguish the truncated from the untruncated $t_n(V, s_0)$ since this truncation results in an error of magnitude $a_n$.

For the discretization step we cover the compact set $A = S(p, q) \times \text{supp}(f_X)$ by finitely many balls, which is possible by (A.4) and the compactness of $S(p, q)$. Let $\delta_n = a_n/h_n$ and $A_j = \{V : ||P_V - P_{V_j}|| \leq \delta_n\} \times \{s : ||s - s_j|| \leq \delta_n\}$ be a cover of $A$ with ball centers $V_j \times s_j$. Then, $A \subset \bigcup_{j=1}^N A_j$ and the number of balls can be bounded by $N \leq C\delta_n^{-d}\delta_n^{-p}$ for some constant $C \in (0, \infty)$, where $d = \dim(S(p, q)) = pq - (q+1)/2$. Let $V \times s_0 \in A_j$. Then by Lemma 10 there exists $0 < C_2 < \infty$, such that

$$|d_i(V, s_0) - d_i(V_j, s_j)| \leq C_2 \delta_n$$

holds for $d_i$ in (15). Under (K.2), which implies (39), inequality (49) yields

$$\left|K\left(\frac{d_i(V, s_0)}{h_n}\right) - K\left(\frac{d_i(V_j, s_j)}{h_n}\right)\right| \leq K^*(\frac{d_i(V_j, s_j)}{h_n})C_2a_n$$

(45)

for $V \times s_0 \in A_j$ and $K^*(\cdot)$ an integrable and bounded function.
Define $r_n^{(l)}(V_j, s_j) = (1/n h_n^{(p-q)/2}) \sum_{i=1}^{n} K^*(d_i(V_j, s_j)/h_n)|Y_i|^l$. For notational convenience we drop the dependence on $l$ and $j$ in the following and observe that (45) yields

$$|t_n^{(l)}(V, s_0) - t_n^{(l)}(V_j, s_j)| \leq C_{22} a_n r_n^{(l)}(V_j, s_j)$$

(46)

Since $K^*$ fulfills (K.1) except for continuity, an analogous argument as in the proof of Lemma 8 yields that $\mathbb{E} \left( r_n^{(l)}(V_j, s_j) \right) < \infty$ by (A.3). By subtracting and adding $t_n^{(l)}(V, s_0)$, $\mathbb{E}(t_n^{(l)}(V_j, s_j))$, the triangular inequality, (46) and integrability of $r_n^{(l)}$, we obtain

$$\left| t_n^{(l)}(V, s_0) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right) \right| \leq \left| t_n^{(l)}(V, s_0) - t_n^{(l)}(V_j, s_j) \right| + \left| \mathbb{E} \left( t_n^{(l)}(V_j, s_j) - t_n^{(l)}(V, s_0) \right) \right|$$

$$+ \left| t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right) \right| \leq C_{22} a_n \left( |r_n| + |\mathbb{E}(r_n)| \right) + \left| t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right) \right|$$

$$\leq 2C_{22} a_n + |r_n - \mathbb{E}(r_n)| + \left| t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right) \right|$$

(47)

for any $C_3 > C_{22} \mathbb{E}(r_n^{(l)}(V_j, s_j))$ and $n$ such that $a_n \leq 1$, since $a_n^2 = o(1)$. Summarizing there exists $0 < C_3 < \infty$ such that (47) holds.

Then using $\sup_{x \in A} f(x) = \max_{1 \leq j \leq N} \sup_{x \in A_j} f(x) \leq \sum_{j=1}^{N} \sup_{x \in A_j} f(x)$ for any partition of $A$ and continuous function $f$, subadditivity of the probability for the first inequality and (47) for the third inequality below, it holds

$$\mathbb{P} \left( \sup_{V \times s_0 \in A} \left| t_n^{(l)}(V, s_0) - \mathbb{E} \left( t_n^{(l)}(V, s_0) \right) \right| > 3C_{3} a_n \right) \leq \sum_{j=1}^{N} \mathbb{P} \left( \sup_{V \times s_0 \in A_j} \left| t_n^{(l)}(V, s_0) - \mathbb{E} \left( t_n^{(l)}(V, s_0) \right) \right| > 3C_{3} a_n \right)$$

$$\leq N \max_{1 \leq j \leq N} \mathbb{P} \left( \sup_{V \times s_0 \in A_j} \left| t_n^{(l)}(V, s_0) - \mathbb{E} \left( t_n^{(l)}(V, s_0) \right) \right| > 3C_{3} a_n \right)$$

$$\leq N \left( \max_{1 \leq j \leq N} \mathbb{P} \left( |t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right)| > C_3 a_n \right) + \max_{1 \leq j \leq N} \mathbb{P} \left( |r_n - \mathbb{E}(r_n)| > C_3 a_n \right) \right) \leq C \delta^{-(d+p)} \left( \max_{1 \leq j \leq N} \mathbb{P} \left( |t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right)| > C_3 a_n \right) + \max_{1 \leq j \leq N} \mathbb{P} \left( |r_n - \mathbb{E}(r_n)| > C_3 a_n \right) \right)$$

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

Finally, we bound the first and second term in the last line of (48) by the Bernstein inequality (38). For the first term in the last line of (48), let $Z_i = Y_i h_n^{(p-q)/2} l_n^{(l)}(V_j, s_j)$, then the $Z_i$ are independent, $|Z_i| \leq b = M_1 r_n = M_1 / a_n$ by (K.1) and the truncation step. For $V_n = \mathbb{V} \mathbb{A}(S_n)$, Lemma 9 yields $nh_n^{(p-q)/2} C \geq V_n$ with $C = b_{2l,2} + \delta$, and set $t = C_3 a_n nh_n^{(p-q)/2}$. The Bernstein inequality (38) yields

$$\mathbb{P} \left( \left| t_n^{(l)}(V_j, s_j) - \mathbb{E} \left( t_n^{(l)}(V_j, s_j) \right) \right| > C_3 a_n \right) < 2 \exp \left( \frac{-t^2/2}{V_n + bt/3} \right)$$
\[
2 \exp \left( -\frac{(1/2)C_3^2a_n^2n^2h_n^{(p-q)}}{n(h_n^{(p-q)})/2+C + (1/3)M_1\tau_nC_3a_nh_n^{(p-q)/2}} \right) \leq 2 \exp \left( -\frac{(1/2)C_3 \log(n)}{C/C_3 + (M_1/3)} \right) = 2n^{-\gamma(C_3)}
\]

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

For the second term in the last line of (48), let \(Z_1 = Y_t^l K^*(d_i(V_j,s_j)/h_n)\) in (38) and proceed analogously to obtain
\[
\mathbb{P} \left( \left| r_n^{(l)}(V_j,s_j) - \mathbb{E} \left( r_n^{(l)}(V_j,s_j) \right) \right| > C_3a_n \right) < 2n^{-\gamma(C_3)}
\]

By (H.1), \(h_n^{(p-q)/2} \leq 1\) for \(n\) large, so that \(\delta_n^{-1} = (a_n h_n)^{-1} \leq n^{-1/2} h_n^{-1} h_n^{(p-q)/4} \leq n^{5/2}\). Further (H.2) implies \(1/(n h_n^{(p-q)/2}) \leq 1\) for \(n\) large, therefore \(\delta_n^{-1} \leq n^{2/(q-p)} \leq n^2\) since \(p-q \geq 1\). Therefore, (48) is smaller than \(4 C_3 \delta_n^{-1}(d+p) - 2/2 - \gamma(C_3)\). For \(C_3\) large enough, we have \(5(d+p)/2 - 2 - \gamma(C_3) < 0\) and \(n^{5(d+p)/2} - 2 - \gamma(C_3) \to 0\). This completes the proof. \(\square\)

**Theorem 12.** Under (A.1), (A.2) and (A.4), (H.1), (K.1), and \(\int_{\mathbb{R}^{p-q}} K(||r_2||^2)dr_2 = 1\),
\[
\sup_{V \times s_0 \in A} \left| t_n^{(l)}(V,s_0) + \chi_{(l=2)} \eta^2 t_0^{(0)}(V,s_0) - \mathbb{E} \left( t_n^{(l)}(V,s_0) \right) \right| = O(h_n), \quad l = 0, 1, 2 \quad (49)
\]

**Proof of Theorem 12.** Let \(\tilde{g}(r_1, r_2) = g(B^T s_0 + B^T Vr_1 + B^T Ur_2)^l f_X(s_0 + Vr_1 + Ur_2)\) where \(r_1, r_2\) satisfy the orthogonal decomposition (5). Then
\[
\mathbb{E} \left( t_n^{(l)}(V,s_0) \right) = \int_{\mathbb{R}^{p-q}} K(||r_2||^2) \int_{\mathbb{R}^p} \tilde{g}(r_1, h_n^{1/2}r_2)dr_1 dr_2 + \chi_{(l=2)} \eta^2 \mathbb{E} \left( t_n^{(0)}(V, s_0) \right) \quad (50)
\]
holds by Lemma 8 for \(l = 0, 1\). For \(l = 2\), \(Y_t^2 = g_2^2 + 2g_1\xi_1 + \xi_1^2\) with \(g_1 = g(B^TX_1)\) and can be handled as in the case of \(l = 0, 1, 2\).

Plugging in (50) the second order Taylor expansion for some \(\xi\) in the neighborhood of 0, \(\tilde{g}(r_1, h_n^{1/2}r_2) = \tilde{g}(r_1, 0) + h_n^{1/2} \nabla_{r_2} \tilde{g}(r_1, 0) r_2 + h_n r_2^2 \nabla_{r_2}^2 \tilde{g}(r_1, \xi) r_2, r_2,\) yields
\[
\mathbb{E} \left( t_n^{(l)}(V,s_0) \right) = \int_{\mathbb{R}^p} \tilde{g}(r_1, 0)dr_1 + \sqrt{h_n} \left( \int_{\mathbb{R}^{p-q}} \nabla_{r_2} \tilde{g}(r_1, 0)dr_2 \right)^T \int_{\mathbb{R}^{p-q}} K(||r_2||^2) r_2dr_2 + h_n \frac{1}{2} \int_{\mathbb{R}^{p-q}} K(||r_2||^2) \int_{\mathbb{R}^p} r_2^T \nabla_{r_2}^2 \tilde{g}(r_1, \xi) r_2dr_1 dr_2 = \int_{\mathbb{R}^p} \tilde{g}(r_1, 0)dr_1 + \sqrt{h_n} \left( \int_{\mathbb{R}^{p-q}} \nabla_{r_2} \tilde{g}(r_1, 0)dr_2 \right)^T \int_{\mathbb{R}^{p-q}} K(||r_2||^2) r_2dr_2 + h_n \frac{1}{2} R(V, s_0)
\]
since \(\int_{\mathbb{R}^p} \tilde{g}(r_1, 0)dr_1 = \mathbb{E} t_n^{(l)}(V, s_0)\) and \(\int_{\mathbb{R}^{p-q}} K(||r_2||^2) r_2dr_2 = 0 \in \mathbb{R}^{p-q}\) due to \(K(|| \cdot ||^2)\) being even. Let \(R(V, s_0) = \int_{\mathbb{R}^{p-q}} K(||r_2||^2) \int_{\mathbb{R}^p} r_2^T \nabla_{r_2} \tilde{g}(r_1, \xi) r_2dr_1 dr_2\). By (A.4) and (A.2) it holds \(||r_2^T \nabla_{r_2} \tilde{g}(r_1, \xi)|| \leq C ||r_2||^2\) for \(C = \sup_{x,y} ||\nabla_{r_2} \tilde{g}(x,y)|| < \infty\), since a continuous function over a compact set is bounded. Then, \(R(V, s_0) \leq C \int_{\mathbb{R}^{p-q}} K(||r_2||^2) ||r_2||^2 dr_2 < \infty\) for some \(C_4 > 0\) since the integral over \(r_1\) is over a compact set by (A.4). \(\square\)

Lemma 13 follows directly from Theorems 11 and 12 and the triangle inequality.
Lemma 13. Suppose (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1) hold. If \(a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)\), and \(a_n/h_n^{(p-q)/2} = O(1)\), then for \(l = 0, 1, 2\)

\[
\sup_{\mathbf{V} \times s_0 \in A} \left| t_l^{(l)}(\mathbf{V}, s_0) + 1_{\{l=2\}} \eta^2 t_0^{(0)}(\mathbf{V}, s_0) - t_{l_n}^{(l)}(\mathbf{V}, s_0) \right| = O_P(a_n + h_n)
\]

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

Theorem 14. Suppose (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1) hold. Let \(a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)\), \(a_n/h_n^{(p-q)/2} = O(1)\), \(\delta_n = \inf_{\mathbf{V} \times s_0 \in A_n} t_0^{(0)}(\mathbf{V}, s_0)\), where \(t_0^{(0)}(\mathbf{V}, s_0)\) is defined in (11), and \(A_n = S(p, q) \times \{x \in \text{supp}(f_X) : |x - \delta \text{supp}(f_X)| \geq b_n\}\), where \(\partial C\) denotes the boundary of the set \(C\) and \(|x - C| = \inf_{x \in C} |x - r|\), for a sequence \(b_n \to 0\) so that \(\delta_n^{-1}(a_n + h_n) \to 0\) for any bandwidth \(h_n\) that satisfies the assumptions. Then,

\[
\sup_{\mathbf{V} \times s_0 \in A} \left| \tilde{g}_l^{(l)}(\mathbf{V}, s_0) - \mu_l(\mathbf{V}, s_0) - 1_{\{l=2\}} \eta^2 t_0^{(0)}(\mathbf{V}, s_0) \right| = O_P(\delta_n^{-1}(a_n + h_n)), \quad l = 0, 1, 2
\]

and

\[
\sup_{\mathbf{V} \times s_0 \in A} \left| \tilde{L}_n(\mathbf{V}, s_0) - \tilde{L}(\mathbf{V}, s_0) \right| = O_P(\delta_n^{-1}(a_n + h_n)) \tag{51}
\]

where \(\tilde{g}_l^{(l)}(\mathbf{V}, s_0), \mu_l(\mathbf{V}, s_0), \tilde{L}_n(\mathbf{V}, s_0)\) and \(\tilde{L}(\mathbf{V}, s_0)\) are defined in (37), (10), (17) and (9), respectively.

Proof of Theorem 14.

\[
\tilde{g}_l^{(l)}(\mathbf{V}, s_0) = \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} = \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)/t_0^{(0)}(\mathbf{V}, s_0)}
\]

We consider the numerator and denominator separately. By Lemma 13

\[
\sup_{\mathbf{V} \times s_0 \in A_n} \left| \frac{t_0^{(0)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} - 1 \right| = \sup_{A_n} \frac{1}{\inf_{A_n} t_0^{(0)}(\mathbf{V}, s_0)} = O_P(\delta_n^{-1}(a_n + h_n))
\]

Next

\[
\sup_{\mathbf{V} \times s_0 \in A_n} \left| \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} - \mu_l(\mathbf{V}, s_0) \right| = \sup_{A_n} \frac{1}{\inf_{A_n} t_0^{(0)}(\mathbf{V}, s_0)} = O_P(\delta_n^{-1}(a_n + h_n)).
\]

Therefore by \(A_n \uparrow A = S(p, q) \times \text{supp}(f_X)\) we get

\[
\lim_{n \to \infty} \sup_{\mathbf{V} \times s_0 \in A_n} \left| \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} - \mu_l(\mathbf{V}, s_0) \right| = \lim_{n \to \infty} \sup_{\mathbf{V} \times s_0 \in A} \left| \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} - \mu_l(\mathbf{V}, s_0) \right|
\]

and in total we obtain

\[
\tilde{g}_l^{(l)}(\mathbf{V}, s_0) = \frac{t_l^{(l)}(\mathbf{V}, s_0)}{t_0^{(0)}(\mathbf{V}, s_0)} = \mu_l + O_P(\delta_n^{-1}(a_n + h_n)) = \mu_l + O_P(\delta_n^{-1}(a_n + h_n)).
\]
For $l = 2$, $Y_i^2 = g(B^TX_i)^2 + 2g(B^TX_i)e_i + e_i^2$, and (51) follows from (9).

\[ \square \]

**Lemma 15.** Under (A.1), (A.2), (A.4), there exists $0 < C_5 < \infty$ such that
\[ |\mu_l(V, s_0) - \mu_l(V_j, s_0)| \leq C_5\|P_V - P_{V_j}\| \]  
for all $s_0 \in \text{supp}(f_X)$

**Proof.** From the representation $\tilde{t}^{(l)}(P_V, s_0)$ in (60) instead of $t^{(l)}(V, s_0)$, we consider $\mu_l(V, s_0) = \mu_l(P_V, s_0)$ as a function on the Grassmann manifold. Then,
\[
\left| \mu_l(P_V, s_0) - \mu_l(P_{V_j}, s_0) \right| = \left| \frac{\tilde{t}^{(l)}(P_V, s_0) - \tilde{t}^{(l)}(P_{V_j}, s_0)}{\tilde{t}^{(l)}(P_V, s_0) - \tilde{t}^{(l)}(P_{V_j}, s_0)} \right|
\]
\[
\leq \frac{\sup_{l} |\tilde{t}^{(l)}(P_V, s_0)|}{\inf_{l} |\tilde{t}^{(l)}(P_V, s_0)|} \left| \tilde{t}^{(l)}(P_V, s_0) - \tilde{t}^{(l)}(P_{V_j}, s_0) \right|
\]
\[
+ \frac{\sup_{l} |\tilde{t}^{(l)}(P_V, s_0)|}{\inf_{l} |\tilde{t}^{(l)}(P_V, s_0)|} \left| \tilde{t}^{(l)}(P_V, s_0) - \tilde{t}^{(l)}(P_{V_j}, s_0) \right|
\]  
with $\sup_{l} |\tilde{t}^{(l)}(P_V, s_0)| < \infty$ and $\inf_{l} |\tilde{t}^{(l)}(P_V, s_0)| = 0$, since $\tilde{t}^{(l)}$ is continuous, $\Sigma > 0$ and $s_0 \in \text{supp}(f_X)$.

By (A.2), $\tilde{g}(x) = g(B^TX)\tilde{f}_X(x)$ is twice continus differentiable and therefore Lipschitz continuous on compact sets. We denote its Lipschitz constant by $\Sigma_{\tilde{f}}$. Therefore,
\[
\left| \tilde{t}^{(l)}(P_V, s_0) - \tilde{t}^{(l)}(P_{V_j}, s_0) \right| \leq \int_{\text{supp}(f_X)} |\tilde{g}(s_0 + P_Vr) - \tilde{g}(s_0 + P_{V_j}r)| \, dr
\]
\[
\leq L \int_{\text{supp}(f_X)} \|P_V - P_{V_j}\| \, dr \leq L \left( \int_{\text{supp}(f_X)} \|r\| \, dr \right) \|P_V - P_{V_j}\|
\]  
where the last inequality is due to the sub-multiplicativity of the Frobenius norm and the integral being finite by (A.4). Plugging (54) in (53) and collecting all constants into $C_5$ yields (52). \[ \square \]

**Proof of Theorem 3.** By (18) and (7),
\[
|L_n(V) - L(V)| \leq \frac{1}{n} \sum_i \left( \hat{L}_n(V, X_i) - \hat{L}(V, X_i) \right) + \frac{1}{n} \sum_i \left( \hat{L}(V, X_i) - E(\hat{L}(V, X)) \right)
\]  
(55)

The first term on the right hand side of (55) goes to 0 in probability uniformly in $V$ by Theorem 14,
\[
\frac{1}{n} \sum_i \hat{L}_n(V, X_i) - \hat{L}(V, X_i) \leq \sup_{V \times s_0 \in A} |\hat{L}_n(V, s_0) - \hat{L}(V, s_0)| = O_P(\delta_n^{-1}(a_n + h_n))
\]  
(56)

The second term in (55) converges to 0 almost surely for all $V \in S(p, q)$ by the strong law of large numbers. In order to show uniform convergence the same technique as in the proof of Theorem 11 is used. Let $B_j = \{ V \in S(p, q) : \|VV^T - V_jV_j^T\| \leq \delta_n \}$ be a cover of $S(p, q) \subset \bigcup_{j=1}^N B_j$ with $N \leq
By the Fubini-Tornelli Theorem we obtain

\[
\|\mu_t(V, X_i) - \mu_t(V_j, X_i)\| \leq C_5\|P_V - P_{V_j}\|.
\]

(57)

Let \( G_n(V) = \sum_i \tilde{L}(V, X_i)/n \) with \( E(G_n(V)) = L(V) \). Using (57) and following the same steps as in the proof of Theorem 11 we obtain

\[
\begin{align*}
|G_n(V) - L(V)| &\leq |G_n(V) - G_n(V_j)| + |G_n(V_j) - L(V_j)| + |L(V) - L(V_j)| \\
&\leq 2C_6\hat{a}_n + |G_n(V_j) - L(V_j)|
\end{align*}
\]

(58)

for \( V \in B_j \) and some \( C_6 > C_5 \). Inequality (58) leads to

\[
P\left( \sup_{V \in S(p, q)} |G_n(V) - L(V)| > 3C_6\hat{a}_n \right) \leq C N P\left( \sup_{V \in B_j} |G_n(V) - L(V)| > 3C_6\hat{a}_n \right) \leq C n^{d/2} P(|G_n(V_j) - L(V_j)| > C_6\hat{a}_n) \leq C n^{d/2} n^{-\gamma(C_6)} \to 0
\]

(59)

where the last inequality in (59) is obtained by applying (38) with \( Z_i = \tilde{L}(V_j, X_i) \), which is bounded since \( \tilde{L}(. , .) \) is continuous on the compact set \( A \), and \( \gamma(C_6) \) a monotone increasing function of \( C_6 \) that can be made arbitrarily large by choosing \( C_6 \) accordingly. Therefore, \( \sup_{V \in S(p, q)} |L_n(V) - L(V)| \leq OP(\delta_n^{-1}(a_n + h_n + \hat{a}_n)) \) with \( \delta_n = \inf_{V \times s_0 \in A_n} t(0)(V, s_0) \), where \( t(0)(V, s_0) \) is defined in (11), and \( A_n = S(p, q) \times \{ s : f_{\lambda}(s) \geq b_n \} \) for a sequence \( b_n \to 0 \) so that \( \delta_n^{-1}(a_n + h_n) \to 0 \) for any bandwidth \( h_n \) that satisfies the assumptions, which implies (19).

Remark. The objective function in (7) can equivalently be viewed as a function from the Grassmann manifold \( Gr(p, q) \) to \([0, \infty)\). To see this, suppose \( V \in S(p, q) \) is an arbitrary basis of a subspace \( M \in Gr(p, q) \). We can identify \( M \) through the projection \( P_M = VV^T \). By (5) we write \( x = Vr_1 + Ur_2 \). By the Fubini-Tornelli Theorem we obtain

\[
\tilde{\tilde{t}}(l) (P_M, s_0) = \int_{supp(f_X)} g(B^T s_0 + B^T P_M x)^l f_X(s_0 + P_M x) dx
\]

\[
= t(l) (V, s_0) \int_{supp(f_X) \cap \mathbb{R}^{p-q}} dx.
\]

(60)

Therefore \( \tilde{\tilde{t}}(l) (P_M, s_0)/\tilde{t}(0) (P_M, s_0) = t(l)(V, s_0)/t(0)(V, s_0) \) and \( \mu_t(\cdot, s_0) \) in (10) can also be viewed as a function from \( Gr(p, q) \) to \( \mathbb{R} \). If the optimization in (13) is over \( Gr(p, q) \), the objective function (7) has a unique minimum at \( \text{span}\{B\}^\perp \) by Theorem 1.

Proof of Theorem 4. We apply Theorem 4.1.1 of [3] to obtain consistency of the conditional variance estimator. This theorem requires three conditions that guarantee the convergence of the minimizer of a sequence of random functions \( L_n(P_V) \) to the minimizer of the limiting function \( L(P_V) \); i.e., \( P_{\text{span}\{B\}^\perp} = \text{argmin} L_n(P_V) \to P_{\text{span}\{B\}^\perp} = \text{argmin} L(P_V) \) in probability. To apply the theorem three conditions have to be met: (1) The parameter space is compact; (2) \( L_n(V) \) is continuous and a measurable function of the data \( (Y_i, X_i^T)_{i=1,...,n} \); (3) \( L_n(V) \) converges uniformly to \( L(V) \) and \( L(V) \) attains a unique global minimum at \( \text{span}\{B\}^\perp \).
Proof of Theorem 6. The Gaussian kernel $K$ satisfies $\partial_z K(z) = -z K(z)$. From (16) and (17) we have $\hat{L}_n = \hat{y}_2 - \hat{y}_1^2$ where $\hat{y}_l = \sum_i w_i Y_i^l$, $l = 1, 2$. We let $K_j = K(d_j(\mathbf{V}, s_0)/h_n)$, suppress the dependence on $\mathbf{V}$ and $s_0$ and write $w_i = K_i/\sum_j K_j$. Then, $\nabla K_i = (-1/h_n^2)K_id_i\nabla d_i$ and $\nabla w_i = -\left(K_i d_i \nabla d_i (\sum_j K_j) - K_i \sum_j K_j d_j \nabla d_j \right) / (h_n \sum_j K_j)^2$. Next,

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

(61)

Then, $\nabla \hat{L}_n = \nabla \hat{y}_2 - 2\bar{y}_1 \nabla \hat{y}_1$, and inserting $\nabla \hat{y}_l$ from (61) yields $\nabla \hat{L}_n = (-1/h_n^2) \sum_i \left( Y_i^2 - \hat{y}_2 - 2\bar{y}_1 (Y_i - \bar{y}_1) \right) w_i d_i \nabla d_i = (1/h_n^2) \left( \hat{L}_n - (Y_i - \bar{y}_1)^2 \right) w_i d_i \nabla d_i$, since $Y_i^2 - \hat{y}_2 - 2\bar{y}_1 (Y_i - \bar{y}_1) = (Y_i - \bar{y}_1)^2 - \hat{L}_n$. 

\[\square\]