Electronic Journal of Statistics
Vol. 16 (2022) 1595–1634
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS1994

The ensemble conditional variance estimator for sufficient dimension reduction*

Lukas Fertl[1] **and Efstathia Bura**[2]

[1] ***d-fine*** *Austria GmbH, Vienna, Austria*
e-mail: Lukas.Fertl@d-fine.at

[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

Abstract: *Ensemble Conditional Variance Estimation* (ECVE) is a novel sufficient dimension reduction (SDR) method in regressions with continuous response and predictors. ECVE applies to general non-additive error regression models and operates under the assumption that the predictors can be replaced by a lower dimensional projection without loss of information. It is a semiparametric forward regression model-based exhaustive sufficient dimension reduction estimation method that is shown to be consistent under mild assumptions. ECVE outperforms *central subspace mean average variance estimation* (csMAVE), its main competitor, under several simulation settings and in a benchmark data set analysis.

Keywords and phrases: Regression, semi-parametric, linear sufficient reduction, central subspace, ensembles.

Received March 2021.

1. Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, Y a univariate continuous response and $\mathbf{X}$ a p-variate continuous predictor, jointly distributed, with $(Y, \mathbf{X}^T)^T : \Omega \to \mathbb{R}^{p+1}$. We consider the linear sufficient dimension reduction model

$$Y = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon), \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^p$ is independent of the random variable ϵ, $\mathbf{B}$ is a $p \times k$ matrix of rank k, and $g_{cs} : \mathbb{R}^{k+1} \to \mathbb{R}$ is an unknown non-constant function, where the index cs stands for central subspace and serves to distinguish (1.1) from other link functions in the sequel.

[29, Thm. 1] showed that if $(Y, \mathbf{X}^T)^T$ has a joint continuous distribution, (1.1) is equivalent to

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}, \tag{1.2}$$

*L. Fertl was supported by Austrian Science Fund (FWF P 30690-N35). E. Bura was supported by Austrian Science Fund (FWF P 30690-N35) and Vienna Science and Technology Fund (WWTF ICT19-018)

where the symbol $\perp\!\!\!\perp$ indicates stochastic independence. The matrix $\mathbf{B}$ is not unique and can be replaced by any basis of its column space, span$\{\mathbf{B}\}$. Let $\mathcal{S}$ denote a subspace of $\mathbb{R}^p$, and let $\mathbf{P}_{\mathcal{S}}$ denote the orthogonal projection onto $\mathcal{S}$ with respect to the usual inner product. If the response Y and predictor vector $\mathbf{X}$ are independent conditionally on $\mathbf{P}_{\mathcal{S}}\mathbf{X}$, then $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ can replace $\mathbf{X}$ as the predictor in the regression of Y on $\mathbf{X}$ without loss of information. Such $\mathcal{S}$'s are called dimension reduction subspaces and their intersection, provided it satisfies the conditional independence condition (1.2), is called the central subspace and denoted by $\mathcal{S}_{Y|\mathbf{X}}$ [see [6, p. 105], [7]].

By their equivalence, under both models (1.1) and (1.2), $Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T\mathbf{X}$ and $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$. Since the conditional distribution of $Y \mid \mathbf{X}$ is the same as that of $Y \mid \mathbf{B}^T\mathbf{X}$, $\mathbf{B}^T\mathbf{X}$ contains all the information in $\mathbf{X}$ for modeling the target variable Y, and can replace $\mathbf{X}$ without any loss of information.

If the error term in model (1.1) is additive with $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$, (1.1) reduces to $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$. Now, $\mathbb{E}(Y \mid \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{X}) = \mathbb{E}(Y \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$, where $\mathcal{S} = \text{span}\{\mathbf{B}\}$. The mean subspace, denoted by $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, is the intersection of all subspaces $\mathcal{S}$ such that $\mathbb{E}(Y \mid \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$ [8]. In this case, (1.1) becomes the classic mean subspace model with $\text{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. [8] showed that the mean subspace is a subset of the central subspace, $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$.

Several *linear sufficient dimension reduction* (SDR) methods estimate $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ consistently ([1, 22, 20, 26]). *Linear* refers to the reduction being a linear transformation of the predictor vector. *Minimum Average Variance Estimation* (`MAVE`) [26] has been the most accurate, and thus, competitive among them. `MAVE` differentiates from the majority of SDR methods in that it is not *inverse regression* based such as, for example, the widely used *Sliced Inverse Regression* (SIR, [21]). `MAVE` requires minimal assumptions on the distribution of $(Y, \mathbf{X}^T)^T$ and is based on estimating the gradients of the regression function $E(Y \mid \mathbf{X})$ via local-linear smoothing [5].

The *central subspace mean average variance estimation* (`csMAVE`) [25, 13] is a `MAVE` extension that consistently and exhaustively estimates the span$\{\mathbf{B}\}$ in model (1.1) without restrictive assumptions limiting its applicability. `csMAVE` has remained largely uncontested since it was proposed by [25]. It is based on repeatedly applying `MAVE` on the sliced target variables $f_u(Y) = 1_{\{s_{u-1} < Y \leq s_u\}}$, for $s_1 < \ldots < s_H$, where 1_A denotes the indicator function of the set A. [25] showed that the mean subspaces of the sliced Y can be combined to recover the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

Several papers made contributions in establishing a road path from the mean to the central subspace [see [27] for a list of references]. [27] recognized that these approaches pointed to the same direction: if one can estimate the mean subspace of $\mathbb{E}(f(\mathbf{X}) \mid Y)$ for sufficiently many functions $f \in \mathcal{F}$ for a family of functions $\mathcal{F}$, then one can recover the central subspace. Such families that are rich enough to obtain the desired outcome are called *characterizing ensembles* by [27], who also proposed and studied them [see [20] for an overview].

In this paper, we extend the *conditional variance estimator* (`CVE`) [10] to the exhaustive *ensemble conditional variance estimator* (`ECVE`) for recovering fully the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ in model (1.1). *Conditional variance estimation* is a

semi-parametric method for the consistent estimation of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ under minimal regularity assumptions on the distribution of $(Y, \mathbf{X}^T)^T$. In contrast to other SDR approaches, it identifies the orthogonal complement of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. Its generalization `ECVE` builds on the ensemble device of [27], adapting and extending ideas from conditional variance estimation [10], to exhaustively estimate the sufficient dimension reduction in (1.1). Here we lift the `CVE` requirement of an additive independent error regression model and extend it to regressions with errors that can depend on the predictors.

In [10], `CVE` was shown to be a consistent estimator for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ in the model $Y = \mathbb{E}(Y \mid \mathbf{P}_{\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}}\mathbf{X}) + \tilde{\epsilon}$ with $\mathbf{X} \perp\!\!\!\perp \tilde{\epsilon}$, which restricts the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ to agree with the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. Here we show `CVE` to be consistent in models with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$. This allows the identification of the mean subspace $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ across regression models, in particular for transformed responses $f_t(Y)$, where f_t are elements of an *ensemble* $\mathcal{F} = \{f_t : t \in \Omega_T\}$. We then combine them to form the consistent `ECVE` estimate of the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ in much richer models of the form (1.1), where the central subspace may be a proper superset of the mean subspace, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subsetneq \mathcal{S}_{Y|\mathbf{X}}$.

The paper is organized as follows. In Section 2 we define the notation and concepts we use throughout the paper. A short motivational example for the *conditional variance estimator* (`CVE`, [10]) is given in Section 2.1. In Section 2.2, *ensembles* are introduced. The *ensemble conditional variance estimator* (ECVE) is defined in Section 3 and its estimation procedure in Section 4. The consistency of the *ensemble conditional variance estimator* for the central subspace is shown in Section 5. `ECVE` is seen to perform better or on a par with `csMAVE`, as well as several refined ensemble MAVE estimators [27, 20], via simulations in Section 6, and in the Boston Housing data in Section 7. We conclude in Section 8.

2. Preliminaries

We denote by $F_{\mathbf{Z}}$ the cumulative distribution function (cdf) of a random variable or vector $\mathbf{Z}$. We drop the subscript, when the attribution is clear from the context. For a vector $\boldsymbol{a}$ and a matrix $\mathbf{A}$, $\|\boldsymbol{a}\|$ and $\|\mathbf{A}\|$ denote the Euclidean and the Frobenius norm, respectively. Scalar product is the usual Euclidean inner product and $\perp$ denotes orthogonality with respect to it. The probability density function of $\mathbf{X}$ is denoted by $f_{\mathbf{X}}$, and its support by $\mathrm{supp}(f_{\mathbf{X}})$. The notation $Y \perp\!\!\!\perp \mathbf{X}$ signifies stochastic independence of the random vector $\mathbf{X}$ and random variable Y. The j-th standard basis vector with zeroes everywhere except for 1 on the j-th position is denoted by $\mathbf{e}_j \in \mathbb{R}^p$, $\boldsymbol{\iota}_p = (1, 1, \ldots, 1)^T \in \mathbb{R}^p$, and $\mathbf{I}_p = (\mathbf{e}_1, \ldots, \mathbf{e}_p)$ is the identity matrix of order p. For any matrix $\mathbf{M} \in \mathbb{R}^{p\times q}$, $\mathbf{P}_{\mathbf{M}}$ denotes the orthogonal projection matrix on its column or range space $\mathrm{span}\{\mathbf{M}\}$; i.e., $\mathbf{P}_{\mathbf{M}} = \mathbf{P}_{\mathrm{span}\{\mathbf{M}\}} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \in \mathbb{R}^{p\times p}$.

For $q \le p$,

$$\mathcal{S}(p, q) = \{\mathbf{V} \in \mathbb{R}^{p\times q} : \mathbf{V}^T\mathbf{V} = \mathbf{I}_q\}, \tag{2.1}$$

denotes the Stiefel manifold that comprizes of all $p\times q$ matrices with orthonormal columns. $\mathcal{S}(p, q)$ is compact with $\dim(\mathcal{S}(p, q)) = pq - q(q+1)/2$ [see [4] and [24,

Sec. 2.1]]. The set

$$Gr(p,q) = \mathcal{S}(p,q)/\mathcal{S}(q,q) \tag{2.2}$$

denotes a Grassmann manifold [12] that contains all q-dimensional subspaces in $\mathbb{R}^p$. $Gr(p,q)$ is the quotient space of $\mathcal{S}(p,q)$ with all $q \times q$ orthonormal matrices in $\mathcal{S}(q,q)$.

2.1. CVE

To motivate the development of the *ensemble conditional variance estimator* (ECVE) we start this section by a simple example describing the *conditional variance estimator* (CVE) [10], which estimates the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ in regressions with additive error term, i.e. $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$.

To appreciate the geometric intuition behind `CVE`, we consider a bivariate standard normal predictor vector, $\mathbf{X} = (X_1, X_2)^T \sim N(\mathbf{0}, \mathbf{I}_2)$, and generate the response from $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon = X_1 + \epsilon$, with $\epsilon \sim N(0, \eta^2)$ independent of $\mathbf{X}$ with $\eta = 0.1$. In this setting, $k = 1$, $g_{cs}(z,u) = z + u \in \mathbb{R}$ and $\mathbf{B} = (1,0)^T$ in model (1.1) is aligned with the first coordinate axis. Since ϵ is additive and independent from $\mathbf{X}$, the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ is equal to the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, i.e. $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$, and `CVE`, which corresponds to `ECVE` with the ensemble containing only the identity function in this case, recovers the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

We see how `CVE` works in Figure 1, where we plot 100 draws of X_1 versus X_2. 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 $\mathbf{B}$ (x-axis), the color has high variation, whereas in the orthogonal direction $(0,1)$ (y-axis), the color has low variation solely due to the error term ϵ. As it is easier to detect patterns in directions of low variability, CVE identifies $\mathbf{B}$ through its orthogonal complement by finding the directions in which the response Y varies the least (right panel) as $\mathbf{X}$ ranges in an affine subspace. The same intuition underlies `ECVE`, even though it is more difficult to visually capture as the variation may occur in higher moments of the conditional distribution of the response given the predictors.

2.2. Ensembles

[27] introduced *ensembles* as a device to extend mean subspace to central subspace SDR methods. The *ensemble* approach of combining mean subspaces in order to recover the central subspace comprises of two components: (a) a rich family of functions of transformations for the response and (b) a sampling mechanism for drawing the functions from the ensemble to ascertain coverage of the central subspace. To distinguish between families of functions and ensembles, [27] use the term *parametric* ensemble. Here, we drop the denomination and call these families *ensembles*.

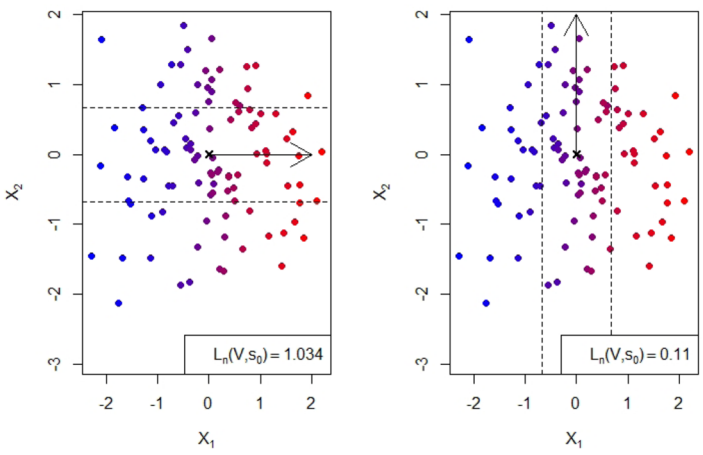

FIG 1. *Plot of* $\mathbf{X}_i$ *samples from* $Y = X_1 + \epsilon$, $i = 1, \ldots, 100$. *The color of the points is determined by their corresponding* Y_i *values, i.e. low* Y_i *values are assigned blue and the higher the* Y_i *value the more red the points are. In the left panel* $\mathbf{V} = \mathbf{B} = (1, 0)^T$, *and in the right panel* $\mathbf{V} = (0, 1)^T \perp \mathbf{B}$, *both with shift point* $\mathbf{s}_0 = (0, 0)^T$ *denoted as black* $\times$. *The subspace* $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ *is indicated with the black arrow and the black dashed lines delineate the slice* $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0 + \text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$.

Definition. *An ensemble is defined to be a family of measurable functions* $\mathcal{F} = \{f_t : t \in \Omega_T\}$ *with respect to an index set* Ω_T *in a Euclidean space.*

Let Y follow model (1.1), $\mathcal{F}$ be an ensemble, and $f \in \mathcal{F}$. The space $\mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}$ is defined to be the mean subspace of the transformed random variable $f(Y)$ [see [6] or [8]].

Definition. *An ensemble* $\mathcal{F}$ *characterizes the central subspace* $\mathcal{S}_{Y|\mathbf{X}}$, *if*

$$\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\} = \mathcal{S}_{Y|\mathbf{X}} \tag{2.3}$$

As an example, the ensemble $\mathcal{F} = \{f_t : t \in \Omega_T\} = \{1_{\{z \leq t\}} : t \in \mathbb{R}\}$ can characterize the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. That is, $\mathbb{E}(f_t(Y)|\mathbf{X})$ is the conditional cumulative distribution function evaluated at t. To see this, let $\mathbf{B} \in \mathcal{S}(p, k)$ be such that $\mathbb{E}(f_t(Y) \mid \mathbf{X}) = \mathbb{E}(f_t(Y) \mid \mathbf{B}^T\mathbf{X})$ for all t. Then, $F_{Y|\mathbf{X}}(t) = \mathbb{E}(f_t(Y) \mid \mathbf{X}) = \mathbb{E}(f_t(Y) \mid \mathbf{B}^T\mathbf{X}) = F_{Y|\mathbf{B}^T\mathbf{X}}(t)$ for all t. Varying over the ensemble $\mathcal{F}$, in this case over $t \in \mathbb{R}$, obtains the conditional cumulative distribution function. This `indicator` ensemble fully recovers the conditional distribution of $Y \mid \mathbf{X}$ and, thus, also the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$,

$$\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\} = \text{span}\{\mathcal{S}_{\mathbb{E}(1_{\{Y \leq t\}}|\mathbf{X})} : t \in \mathbb{R}\} = \mathcal{S}_{Y|\mathbf{X}}$$

We reproduce a list of ensembles $\mathcal{F}$ and associated regularity conditions that can characterize $\mathcal{S}_{Y|\mathbf{X}}$ from [27] next.

Characteristic ensemble: $\mathcal{F} = \{f_t : t \in \Omega_T\} = \{\exp(it\cdot) : t \in \mathbb{R}\}$

Indicator ensemble: $\mathcal{F} = \{1_{\{z \le t\}} : t \in \mathbb{R}\}$, where $\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\}$ recovers the conditional cumulative distribution function

Kernel ensemble: $\mathcal{F} = \{h^{-1}K\left((z-t)/h\right) : t \in \mathbb{R}, h > 0\}$, where K is a kernel suitable for density estimation, and $\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\}$ recovers the conditional density

Polynomial ensemble: $\mathcal{F} = \{z^t : t = 1, 2, 3, ...\}$, where $\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\}$ recovers the conditional moment generating function

Box-Cox ensemble: $\mathcal{F} = \{(z^t - 1)/t : t \neq 0\} \cup \{\log(z) : t = 0\}$ Box-Cox Transforms

Wavelet ensemble: Haar Wavelets

The `characteristic` and `indicator` ensembles describe the conditional characteristic and distribution function of $Y \mid \mathbf{X}$, respectively, which always exist and uniquely determine the distribution. If the conditional density function $f_{Y|\mathbf{X}}$ of $Y \mid \mathbf{X}$ exists, then the *kernel* ensemble characterizes the conditional distribution $Y \mid \mathbf{X}$. Further, if the conditional moment generating function exists, then the polynomial ensemble characterizes $\mathcal{S}_{Y|\mathbf{X}}$. [27] used the ensemble device to extend `MAVE` [26], which targets the mean subspace, to its ensemble version that also estimates the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ consistently.

Theorem 2.1 [27, Thm 2.1] establishes when an ensemble $\mathcal{F}$ is rich enough to characterize $\mathcal{S}_{Y|\mathbf{X}}$.

Theorem 2.1. *Let $\mathcal{B} = \{1_A : A$ is a Borel set in supp$(Y)\}$ be the set of indicator functions on supp(Y), and $L^2(F_Y)$ be the set of square integrable random variables with respect to the marginal distribution F_Y. If $\mathcal{F} \subseteq L^2(F_Y)$ is dense in $\mathcal{B} \subseteq L^2(F_Y)$, then the ensemble $\mathcal{F}$ characterizes the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.*

In Theorem 2.2 we show that finitely many functions of an ensemble $\mathcal{F}$ are sufficient to characterize the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$.

Theorem 2.2. *Let $t_1, t_2, \ldots$ be an i.i.d. sequence of random variables on Ω_T. If an ensemble $\mathcal{F}$ characterizes $\mathcal{S}_{Y|\mathbf{X}}$, then the event*

$$\{\omega : \text{there exists a finite integer } m_0(\omega) \text{ such that for all } m > m_0(\omega), \\ \text{span}\{\mathcal{S}_{\mathbb{E}(f_{t_i}(Y)|\mathbf{X})} : i \in 1, \ldots, m\} = \mathcal{S}_{Y|\mathbf{X}}\}$$

has probability 1.

Theorem 2.2 is Theorem 2.2 in [27], where its proof can be found. The importance of Theorem 2.2 lies in the fact that the search to characterize the central subspace is over a finite set with probability 1. Theorem 2.2 does not offer tools for identifying the elements of the ensemble and is not used in any proofs in the paper.

3. Ensemble CVE for the central subspace

Throughout the paper, we refer to the following assumptions as needed.

(E.1). *Model* (1.1), $Y = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon)$, *holds with* $Y \in \mathbb{R}$, $g_{cs} : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ *non constant in the first argument,* $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathcal{S}(p, k)$, $\mathbf{X} \in \mathbb{R}^p$ *independent of* ϵ, *the distribution of* $\mathbf{X}$ *is absolutely continuous with respect to the Lebesgue measure in* $\mathbb{R}^p$, $supp(f_{\mathbf{X}})$ *is convex, and* $\mathbb{V}\mathrm{ar}(\mathbf{X}) = \mathbf{\Sigma}_{\mathbf{x}}$ *is positive definite.*

(E.2). *The density* $f_{\mathbf{X}} : \mathbb{R}^p \to [0, \infty)$ *of* $\mathbf{X}$ *is twice continuously differentiable with compact support* $supp(f_{\mathbf{X}})$.

(E.3). *The index set* Ω_T *of an ensemble* $\mathcal{F}$ *is endowed with a probability measure* F_T *such that for all* $t \in \Omega_T$ *with* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \neq \{\mathbf{0}\}$,

$$\mathbb{P}_{F_T}\left(\{\tilde{t} \in \Omega_T : \mathcal{S}_{\mathbb{E}(f_{\tilde{t}}(Y)|\mathbf{X})} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}\}\right) > 0.$$

(E.4). *For all* $f \in \mathcal{F}$, *where* $\mathcal{F}$ *is an ensemble, the conditional expectation*

$$\mathbb{E}\left(f(Y) \mid \mathbf{X}\right)$$

is twice continuously differentiable in the conditioning argument. Further, for all $f \in \mathcal{F}$

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

Assumption (E.1) assures the existence and uniqueness of $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}\{\mathbf{B}\}$. Furthermore, it allows the mean subspace to be a proper subset of the central subspace, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subsetneq \mathcal{S}_{Y|\mathbf{X}}$. In Assumption (E.2), the compactness requirement for $\mathrm{supp}(f_{\mathbf{X}})$ is not as restrictive as it might seem. [28, Prop. 11] showed that there is a compact set $K \subset \mathbb{R}^p$ such that $\mathcal{S}_{Y|\mathbf{X}_{|K}} = \mathcal{S}_{Y|\mathbf{X}}$, where $\mathbf{X}_{|K} = \mathbf{X}1_{\{\mathbf{X}\in K\}}$. Assumption (E.3) implies that the set of indices that characterize the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ is not a null set. In practice, the choice of the probability measure F_T on the index set Ω_T of an ensemble $\mathcal{F}$ can always guarantee the fulfillment of this assumption. If the characteristic or indicator ensemble are used, (E.4) states that the conditional characteristic or distribution function is twice continuously differentiable. In this case, the $8th$ moment exists since the complex exponential and indicator functions are bounded.

Definition. *Let* $\mathcal{F}$ *be an ensemble and* $f \in \mathcal{F}$. *For* $q \leq p \in \mathbb{N}$, *and any* $\mathbf{V} \in S(p, q)$, *we define*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\mathrm{ar}\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}\right) \tag{3.1}$$

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

Definition. *Let* F_T *be a cumulative distribution function on the index set* Ω_T *of an ensemble* $\mathcal{F}$. *For* $q \leq p$, *and any* $\mathbf{V} \in S(p, q)$, *we define*

$$\begin{aligned} L_{\mathcal{F}}(\mathbf{V}) &= \int_{\Omega_T}\int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}, f_t) dF_{\mathbf{X}}(\mathbf{x}) dF_T(t) \\ &= \mathbb{E}_{t\sim F_T}\left(\mathbb{E}_{\mathbf{X}}\left(\tilde{L}(\mathbf{V}, \mathbf{X}, f_t)\right)\right) = \mathbb{E}_{t\sim F_T}(L^*(\mathbf{V}, f_t)), \end{aligned} \tag{3.2}$$

where $F_{\mathbf{X}}$ *is the cdf of* $\mathbf{X}$,*and*

$$L^*(\mathbf{V}, f_t) = \mathbb{E}_{\mathbf{X}}\left(\tilde{L}(\mathbf{V}, \mathbf{X}, f_t)\right). \tag{3.3}$$

For the identity function, $f_{t_0}(z) = z$, (3.3) is the target function of the *conditional variance estimation* proposed in [10]. If the random variable t is concentrated on t_0; i.e., $t \sim \delta_{t_0}$, then the *ensemble conditional variance estimator* (ECVE) coincides with the *conditional variance estimator* (CVE).

Next we define the *ensemble conditional variance estimator* (ECVE) for an ensemble $\mathcal{F}$ that characterizes the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. Following the *ensemble minimum average variance estimation* formulation in [27], we extend the original objective function by integrating over the index random variable $t \sim F_T$ that indexes $\mathcal{F}$ in (3.2).

Definition 1. *Let*

$$\mathbf{V}_q = \operatorname{argmin}_{\mathbf{V} \in S(p,q)} L_{\mathcal{F}}(\mathbf{V}) \tag{3.4}$$

The ***Ensemble Conditional Variance Estimator*** *with respect to the ensemble $\mathcal{F}$ is defined to be any basis $\mathbf{B}_{p-q,\mathcal{F}}$ of* $\operatorname{span}\{\mathbf{V}_q\}^{\perp}$.

Theorem 3.1. *Assume (E.1), (E.2), (E.3), and (E.4) hold, and that the function $h(\cdot)$ defined in Theorem A.2 in Appendix A is continuous. Let $\mathcal{F}$ be an ensemble that characterizes $\mathcal{S}_{Y|\mathbf{X}}$, with $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$, and $\mathbf{V}$ be an element of the Stiefel manifold $S(p,q)$ with $q = p - k$. Then, $\mathbf{V}_q$ in (3.4) is well defined and*

$$\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{V}_q\}^{\perp}. \tag{3.5}$$

4. Estimation of the ensemble conditional variance estimator

Assume $(Y_i, \mathbf{X}_i^\top)_{i=1,\dots,n}^\top$ is an i.i.d. sample from model (1.1), and let

$$\begin{aligned} d_i(\mathbf{V}, \mathbf{s}_0) &= \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}}\mathbf{X}_i\|_2^2 = \|\mathbf{X}_i - \mathbf{s}_0\|_2^2 - \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{V}\mathbf{V}^\top(\mathbf{X}_i - \mathbf{s}_0)\rangle \\ &= \|(\mathbf{I}_p - \mathbf{V}\mathbf{V}^\top)(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 = \|\mathbf{Q}_{\mathbf{V}}(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 \end{aligned} \tag{4.1}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^\top$ and $\mathbf{Q}_{\mathbf{V}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{V}}$. The estimators we propose involve a variation of kernel smoothing, which depends on a bandwidth h_n. In our procedure, h_n is the squared width of a slice around the affine subspace $\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$. In order to obtain pointwise convergence for the ensemble CVE, we require the following bias and variance assumptions on the bandwidth, as is typical in nonparametric estimation.

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

(H.2). $n h_n^{(p-q)/2} \to \infty$ *as* $n \to \infty$.

In order to obtain consistency of the proposed estimator, Assumption (H.2) will be strengthened to $\log(n)/n h_n^{(p-q)/2} \to 0$.

We also let K, which we refer to as *kernel*, be a function satisfying the following assumptions (K.1) and (K.2).

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

(K.2). *There exist positive finite constants L_1 and L_2 such that K satisfies either (1) or (2) below:*

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

The Gaussian kernel $K(z) = \exp(-z^2)$, for example, fulfills both (K.1) and (K.2) [see [14]], and is used throughout the paper. For $i = 1, \ldots, n$, we let

$$w_i(\mathbf{V}, \mathbf{s}_0) = \frac{K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{d_j(\mathbf{V},\mathbf{s}_0)}{h_n}\right)} \tag{4.2}$$

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0, f) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0) f(Y_i)^l \quad \text{for} \quad l = 1, 2. \tag{4.3}$$

We estimate $\tilde{L}(\mathbf{V}, s_0, f)$ in (A.3) with

$$\tilde{L}_n(\mathbf{V}, s_0, f) = \bar{y}_2(\mathbf{V}, \mathbf{s}_0, f) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0, f)^2, \tag{4.4}$$

and the objective function $L^*(\mathbf{V}, f)$ in (3.3) with

$$L_n^*(\mathbf{V}, f) = \frac{1}{n}\sum_{i=1}^n \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f), \tag{4.5}$$

where each data point $\mathbf{X}_i$ is a shifting point. For an ensemble $\mathcal{F} = \{f_t : t \in \Omega_T\}$ and $(t_j)_{j=1,\ldots,m_n}$ an i.i.d. sample from F_T with $\lim_{n\to\infty} m_n = \infty$, we use

$$L_{n,\mathcal{F}}(\mathbf{V}) = \frac{1}{m_n}\sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) \tag{4.6}$$

to estimate the objective function in (3.2). The *ensemble conditional variance estimator* (ECVE) is defined to be any basis of $\text{span}\{\widehat{\mathbf{V}}_q\}^\perp$, where

$$\widehat{\mathbf{V}}_q = \text{argmin}_{\mathbf{V}\in S(p,q)} L_{n,\mathcal{F}}(\mathbf{V}) \tag{4.7}$$

We use a similar algorithm to that in [10] to solve the optimization problem (4.7). It requires the explicit form of the gradient of (4.6). Theorem 4.1 provides the gradient when a Gaussian kernel is used.

Theorem 4.1. *The gradient of $\tilde{L}_n(\mathbf{V}, s_0, f)$ in (4.4) is given by*

$$\nabla_{\mathbf{V}}\tilde{L}_n(\mathbf{V}, s_0, f) =$$

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

and the gradient of $L_{n,\mathcal{F}}(\mathbf{V})$ *in* (4.6) *is*

$$\nabla_{\mathbf{V}} L_{n,\mathcal{F}}(\mathbf{V}) = \frac{1}{nm_n} \sum_{i=1}^{n} \sum_{j=1}^{m_n} \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f_{t_j}).$$

In the implementation of `ECVE`, we follow [10] and set the bandwidth to

$$h_n = 1.2^2 \, \frac{2\mathrm{tr}(\widehat{\mathbf{\Sigma}}_{\mathbf{x}})}{p} \left(n^{-1/(4+p-q)}\right)^2. \tag{4.8}$$

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

4.1. Weighted estimation of $L_n^*(\mathbf{V}, f)$

The set of points $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\mathrm{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$ represents a *slice* in the subspace of $\mathbb{R}^p$ about $\mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}$. In the estimation of $L(\mathbf{V})$ two different weighting schemes are used: (a) *Within slices*: The weights are defined in (4.2) and are used to calculate (4.4). (b) *Between slices*: Equal weights $(1/n)$ are used to calculate (4.5). Another idea for the between slices weighting is to assign more weight to slices with more points. This can be realized by altering (4.5) to

$$L_n^{(w)}(\mathbf{V}, f) = \sum_{i=1}^{n} \tilde{w}(\mathbf{V}, \mathbf{X}_i) \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f), \quad \text{with} \tag{4.9}$$

$$\begin{aligned} \tilde{w}(\mathbf{V}, \mathbf{X}_i) &= \frac{\sum_{j=1}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n) - 1}{\sum_{l,u=1}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n) - n} \\ &= \frac{\sum_{j=1, j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)}{\sum_{l,u=1, l\neq u}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n)} \end{aligned} \tag{4.10}$$

The denominator in (4.10) guarantees the weights $\tilde{w}(\mathbf{V}, \mathbf{X}_i)$ sum up to one.

If (4.9) instead of (4.5) is used in (4.6) we refer to this method as *weighted ensemble conditional variance estimation*. For example, if a rectangular kernel is used, $\sum_{j=1, j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$ equals the number of $\mathbf{X}_j$ $(j \neq i)$ points in the slice corresponding to $\tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f)$. Therefore, this slice is assigned weight that is proportional to the number of $\mathbf{X}_j$ points in it, and the more observations we use for estimating $L(\mathbf{V}, \mathbf{X}_i, f)$, the better its accuracy.

5. Consistency of `ECVE`

The consistency of `ECVE` derives from the fact that we can recover $\mathcal{S}_{Y|\mathbf{X}}$ from $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ by varying over $f_t \in \mathcal{F} = \{f_t : t \in \Omega_T\}$ for an ensemble that characterizes $\mathcal{S}_{Y|\mathbf{X}}$. This is achieved in sequential steps, starting from Theorem 5.1, which is the main building block, to Theorem 5.4. The proofs are technical and lengthy, and are, thus, given in Appendix B.

Theorem 5.1. *Assume conditions (E.1), (E.2), (E.4), (K.1), (K.2), (H.1) hold,* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$*, and* $a_n/h_n^{(p-q)/2} = O(1)$*. Let* $\mathcal{F}$ *be an ensemble such that* $\mathbb{E}(|\tilde{\epsilon}|^l \mid \mathbf{X} = \mathbf{x})$ *is continuous for* $l = 1, \ldots, 4$*, and the second conditional moment is twice continuously differentiable, where* $\tilde{\epsilon}$ *is given by Theorem A.1 in Appendix A. Then,* $L_n^*(\mathbf{V}, f)$ *in (4.5) converges uniformly in probability to* $L^*(\mathbf{V}, f)$ *in (3.3) for all* $f \in \mathcal{F}$*. That is,*

$$\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f) - L^*(\mathbf{V}, f)| \longrightarrow 0 \quad \textit{in probability as } n \to \infty.$$

Next, Theorem 5.2 shows that the ensemble conditional variance estimator is consistent for $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ for any transformation f_t.

Theorem 5.2. *Under the same conditions as Theorem 5.1, the conditional variance estimator* $\operatorname{span}\{\widehat{\mathbf{B}}_{k_t}^t\}$ *estimates* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ *consistently, for* $f_t \in \mathcal{F}$*. That is,*

$$\|\mathbf{P}_{\widehat{\mathbf{B}}_{k_t}^t} - \mathbf{P}_{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}}\| \to 0 \quad \textit{in probability as } n \to \infty,$$

where $\widehat{\mathbf{B}}_{k_t}^t$ *is any basis of* $\operatorname{span}\{\widehat{\mathbf{V}}_{k_t}^t\}^\perp$ *with*

$$\widehat{\mathbf{V}}_{k_t}^t = \operatorname{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L_n^*(\mathbf{V}, f_t),$$

with $q = p - k_t$ *and* $k_t = \dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})})$*.*

A straightforward application of Theorem 5.2, using the identity function, obtains that $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ can be consistently estimated by `ECVE`.

Theorem 5.3. *Assume the conditions of Theorem 5.1 hold. Let* $\mathcal{F}$ *be an ensemble such that* $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$ *almost surely, and let the index random variable* $t \sim F_T$ *be independent from the data* $(Y_i, \mathbf{X}_i)_{i=1,\ldots,n}$*. Then* $L_{n,\mathcal{F}}(\mathbf{V})$ *in (4.6) converges uniformly in probability to* $L_{\mathcal{F}}(\mathbf{V})$ *in (3.2); i.e.,*

$$\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_{n,\mathcal{F}}(\mathbf{V}) - L_{\mathcal{F}}(\mathbf{V})| \longrightarrow 0 \quad \textit{in probability as } n \to \infty.$$

The assumption $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$ in Theorem 5.3 is trivially satisfied by the elements of the characteristic and indicator ensembles. Further, the assumption $a_n/h_n^{(p-q)/2} = O(1)$ used for the truncation step in the proof of Theorem 5.1 (see Appendix B) can be dropped since obviously no truncation is needed.

The rate of convergence of m_n is not characterized in Theorem 5.3. In the simulation studies of Section 6, we find that m_n should be chosen to be very small relative to the sample size n, roughly at the rate of $\log(n)$. The consistency of `ECVE` is shown in Theorem 5.4.

Theorem 5.4. *Assume the conditions of Theorem 5.1 and (E.3) hold. Let* $\mathcal{F}$ *be an ensemble that characterizes* $\mathcal{S}_{Y|\mathbf{X}}$ *and whose members satisfy* $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$ *almost surely. Also, assume the index random variable* $t \sim F_T$ *is independent from the data* $(Y_i, \mathbf{X}_i)_{i=1,\ldots,n}$*. Then, the ensemble conditional*

variance estimator (ECVE) is a consistent estimator for $\mathcal{S}_{Y|\mathbf{X}}$. *That is, for any basis* $\widehat{\mathbf{B}}_{p-q,\mathcal{F}}$ *of* $\operatorname{span}\{\widehat{\mathbf{V}}_q\}^{\perp}$, *where* $\widehat{\mathbf{V}}_q$ *is defined in* (4.7) *with* $q = p - k$ *and* $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$,

$$\|\mathbf{P}_{\widehat{\mathbf{B}}_{p-q,\mathcal{F}}} - \mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\| \longrightarrow 0 \quad \textit{in probability as } n \to \infty,$$

where $\mathbf{P}_{\mathbf{M}}$ *denotes the orthogonal projection onto the range space of the matrix or linear subspace* $\mathbf{M}$.

6. Simulation studies

6.1. Influence of m_n on ECVE

In Theorem 5.3 and 5.4, how fast m_n approaches ∞ is unspecified. In this section we study the influence of m_n, the number of functions of the ensemble $\mathcal{F}$ in (4.6), on the accuracy of the ensemble conditional variance estimation.

We consider the 2-dimensional regression model

$$Y = (\mathbf{b}_2^T\mathbf{X}) + (0.5 + (\mathbf{b}_1^T\mathbf{X})^2)\epsilon, \tag{6.1}$$

where $p = 10$, $k = 2$, $\mathbf{X} \sim N(0, \mathbf{I}_{10})$, $\epsilon \sim N(0,1)$ independent of $\mathbf{X}$, $\mathbf{b}_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^p$, and $\mathbf{b}_2 = (0, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$. Therefore, $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \operatorname{span}\{\mathbf{b}_2\} \subsetneq \mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$, with $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2) : p \times 2$ of rank 2.

We set the sample size to $n = 300$ and vary m over $\{4, 8, 10, 26, 50, 76, 100\}$ for the (a) indicator, $\mathcal{F}_{m,\text{Indicator}} = \{1_{\{x \geq q_j\}} : j = 1, \ldots, m\}$, where q_j is the $j/(m+1)$th empirical quantile of $(Y_i)_{i=1,\ldots,n}$; (b) characteristic or Fourier, $\mathcal{F}_{m,\text{Fourier}} = \{\sin(jx) : j = 1, \ldots, m/2\} \cup \{\cos(jx) : j = 1, \ldots, m/2\}$; (c) monomial, $\mathcal{F}_{m,\text{Monom}} = \{x^j : j = 1, \ldots, m\}$, (d) and Box-Cox, $\mathcal{F}_{m,\text{BoxCox}} = \{(x^{t_j} - 1)/t_j : t_j = 0.1 + 2(j-1)/(m-1), j = 1, \ldots, m-1\} \cup \{\log(x)\}$, ensembles.

The accuracy of the estimates throughout this paper is assessed using

$$err = \frac{\|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|}{\sqrt{2k}} \in [0, 1], \tag{6.2}$$

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

For each ensemble, we form the ensemble conditional variance estimator and its weighted version as in Section 4.1 (see also [10]). The results of 100 replications for each method and each value of m are displayed in Figure 2. Specifically, in model (6.1), $\text{err}_{j,m} = \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|/(2k)^{1/2}$, $j = 1, \ldots, 100$, $m \in \{2, 4, 8, 10, 26, 50, 76, 100\}$. `ECVE`'s main competitor, `csMAVE` does not vary with m. `csMAVE`'s estimate of the central subspace has median error 0.2 with a wide range from 0.1 to 0.6. The estimation error of Fourier, Indicator and Box-Cox `ECVE` varies over m, and is on par or better than `csMAVE`'s for some m values.

For the Fourier basis, fewer basis functions give the best performance. The indicator and Box-Cox ensembles are quite robust against varying m, whereas the errors get rapidly larger if m is increased for the monomial ensemble. The weighted version of `ECVE` improves the accuracy for all ensembles. $\mathcal{F}_{4,\text{Fourier\_weighted}}$, $\mathcal{F}_{8,\text{Indicator\_weighted}}$, $\mathcal{F}_{4,\text{BoxCox\_weighted}}$ are on par or more accurate than `csMAVE`.

In sum, the simulation results support the choice of a small m number of basis functions. Based on this and further unreported simulations, we set the default value of m to

$$m_n = \begin{cases} \lceil \log(n) \rceil, & \text{if} \quad \lceil \log(n) \rceil \quad \text{even} \\ \lceil \log(n) \rceil + 1, & \text{if} \quad \lceil \log(n) \rceil \quad \text{odd} \end{cases} \tag{6.3}$$

for all simulations in Sections 6.2, 6.3 and the data analysis in Section 7, where $\lceil x \rceil$ is the smallest integer greater than or equal to x.

6.2. Demonstrating consistency

We explore the consistency rate of the *conditional variance estimator* (`CVE`) and *ensemble conditional variance estimator* (`ECVE`), `csMAVE` and `mMAVE` in model (6.1).

Specifically, we apply seven estimation methods, the first five targeting the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ and the last two $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, as follows. For $\mathcal{S}_{Y|\mathbf{X}}$, we compare `ECVE` for the indicator (I), Fourier (II), monomial (III) and Box-Cox (IV) ensembles, as in Section 6.1, and `csMAVE` (V). For $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, we use `CVE` (VI) of [10] and `mMAVE` (VII) in [26].

The simulation is performed as follows. We generate 100 i.i.d samples from (6.1), $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$, for each sample size $n = 100, 200, 400, 600, 800, 1000$. Recall that (6.1) is a two dimensional model with $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}(\mathbf{b}_2) \subsetneq \mathcal{S}_{Y|\mathbf{X}} = \text{span}(\mathbf{B})$. For methods (I)-(V), we set $k = 2$ and estimate $\mathbf{B} \in \mathbb{R}^{10\times 2}$. For (VI) and (VII), we set $k = 1$ and estimate $\mathbf{b}_2 \in \mathbb{R}^{10\times 1}$. Then, we calculate $\text{err}_{j,n} = \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|/(2k)^{1/2}$, $j = 1, \ldots, 100$, $n \in \{100, 200, 400, 600, 800, 1000\}$.

Figure 3 displays the distribution of $err_{j,n}$ for the seven methods. As the sample size increases, `ECVE` Indicator, Fourier and `csMAVE` are on par with respect to both speed and accuracy. The accuracy of `ECVE` Box-Cox improves as the sample size increases but at a slower rate. There is no improvement in the accuracy of `ECVE` monomial. This is not surprising as the monomial, as well as the Box-Cox, do not satisfy the assumption $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$ in Theorem 5.4, in contrast to the Indicator and Fourier ensembles. The Fourier and Indicator `ECVE`, and `csMAVE` estimate $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$ consistently. The mean subspace methods, `CVE` and `mMAVE`, estimate $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{b}_2\}$ consistently.

6.3. Evaluating estimation error

We consider seven models M1-M7, defined in Table 1, three different sample sizes $\{100, 200, 400\}$, and three different distributions of the predictor vector

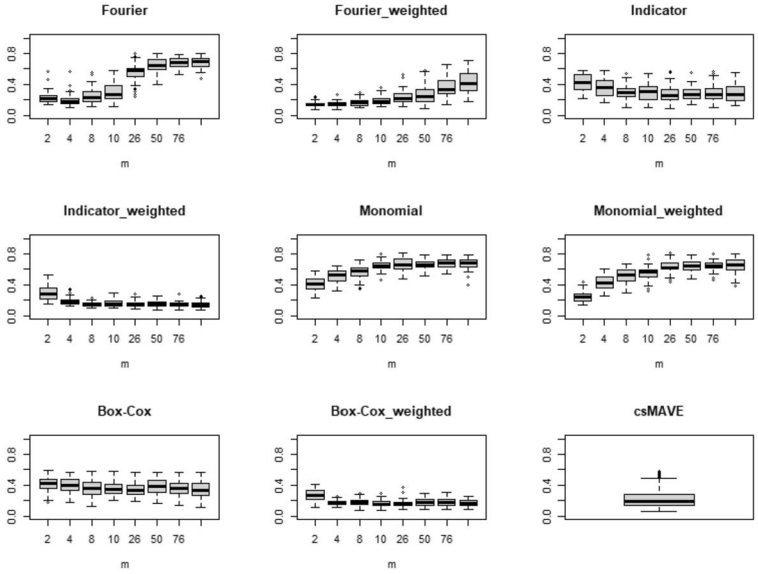

FIG 2. *Box plots of the estimation errors over* 100 *replications of model* (6.1) *with* $n = 300$ *over* $m = |\mathcal{F}| = 2, 4, 8, 10, 26, 50, 76, 100$*, across four ensembles and* `csMAVE`*.*

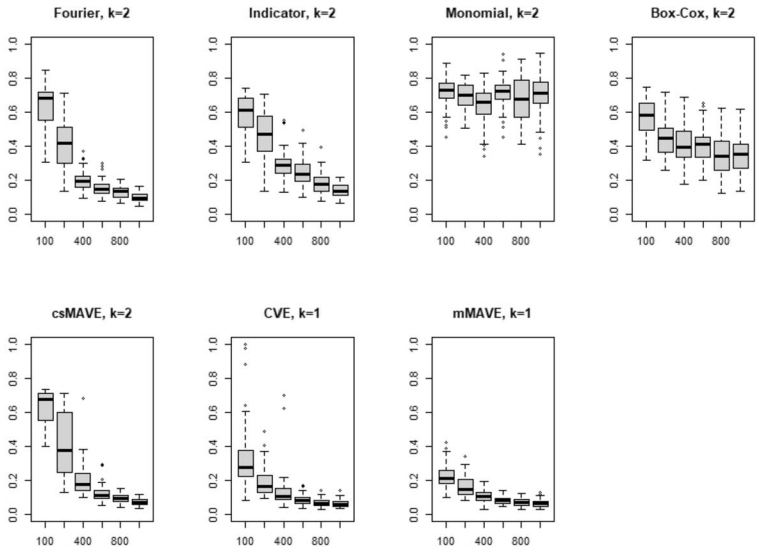

FIG 3. *Box-plots of estimation errors for model* (6.1) *over* $n = 100, 200, 400, 600, 800, 1000$ *for the seven (I-VII) methods in Section 6.2*

$\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Z} \in \mathbb{R}^p$, where $\mathbf{\Sigma} = (\Sigma_{ij})_{i,j=1,\ldots,p}$, $\Sigma_{i,j} = 0.5^{|i-j|}$. Throughout, $p = 10$, $\mathbf{B}$ are the first k columns of $\mathbf{I}_p$, and $\epsilon \sim N(0,1)$ independent of $\mathbf{X}$.

As in [25], we consider three distributions for $\mathbf{Z} \in \mathbb{R}^p$: (I) $N(0, \mathbf{I}_p)$, (II) p-dimensional uniform distribution on $[-\sqrt{3}, \sqrt{3}]^p$; i.e., all components of $\mathbf{Z}$ are independent and uniformly distributed, and (III) a mixture-distribution $N(0, \mathbf{I}_p) + \boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T \in \mathbb{R}^p$ with $\mu_j = 2$, $\mu_k = 0$, for $k \neq j$, and j is uniformly distributed on $\{1, \ldots, p\}$.

The simple and weighted [see Section 4.1] Fourier and indicator ensembles are used to form four *ensemble conditional variance estimators*. The monomial and Box-Cox ensembles were also used but did not give satisfactory results and are not reported. From these two ensembles, four `ECVE` estimators are formed and compared against the reference methods `csMAVE` [25], as implemented in the `R` package `MAVE`, *refined ensemble minimum average variance estimation* (`reMAVE`, [27, 20]), and *refined ensemble outer product gradient* (`reOPG` [27, 20]). The source code for `CVE` and its ensemble version is available at https://git.art-ist.cc/daniel/CVE.

We applied `reMAVE` and `reOPG` with the Fourier, Box-Cox, and monomial ensembles. Only the results of the former two are reported as the Monomial ensemble exhibited non-competitive performance. The same occurred for `reOPG`, whose results are also not reported as it was always outperformed by the corresponding `reMAVE`. For example, `reMAVE` using the Fourier ensemble outperformed `reOPG` using the same ensemble. This is not surprising since `reMAVE` uses `reOPG` as a starting point in its optimization procedure. The source code for `reMAVE` and `reOPG` is taken from [20]. The ensemble size and the number of iterations are both set to 10. In total, we report the results from seven methods: `ECVE` and its weighted version with Fourier ensemble denoted as `Fourier` and `Fourier_weighted`, `ECVE` and its weighted version with Indicator ensemble denoted as `Indicator` and `Indicator_weighted`, `csMAVE` (which corresponds to `reMAVE` with the indicator ensemble), and `reMAVE` with the Fourier and Box-Cox ensembles denoted by `reMAVEf` and `reMAVEb`.

TABLE 1

Name	Model	$\mathcal{S}_{\mathbb{E}(Y\|\mathbf{X})}$	$\mathcal{S}_{Y\|\mathbf{X}}$	k
M1	$Y = \frac{1}{\mathbf{b}_1^T\mathbf{X}} + 0.2\epsilon$	$\text{span}\{\mathbf{b}_1\}$	$\text{span}\{\mathbf{b}_1\}$	1
M2	$Y = \cos(2\mathbf{b}_1^T\mathbf{X}) + \cos(\mathbf{b}_2^T\mathbf{X}) + 0.2\epsilon$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	2
M3	$Y = (\mathbf{b}_2^T\mathbf{X}) + (0.5 + (\mathbf{b}_1^T\mathbf{X})^2)\epsilon$	$\text{span}\{\mathbf{b}_2\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	2
M4	$Y = \frac{\mathbf{b}_1^T\mathbf{X}}{0.5+(1.5+\mathbf{b}_2^T\mathbf{X})^2} + (\lvert\mathbf{b}_1^T\mathbf{X}\rvert + (\mathbf{b}_2^T\mathbf{X})^2 + 0.5)\epsilon$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	2
M5	$Y = \mathbf{b}_3^T\mathbf{X} + \sin(\mathbf{b}_1^T\mathbf{X}(\mathbf{b}_2^T\mathbf{X})^2)\epsilon$	$\text{span}\{\mathbf{b}_3\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$	3
M6	$Y = 0.5(\mathbf{b}_1^T\mathbf{X})^2\epsilon$	$\text{span}\{\mathbf{0}\}$	$\text{span}\{\mathbf{b}_1\}$	1
M7	$Y = \cos(\mathbf{b}_1^T\mathbf{X} - \pi) + \cos(2\mathbf{b}_1^T\mathbf{X})\epsilon$	$\text{span}\{\mathbf{b}_1\}$	$\text{span}\{\mathbf{b}_1\}$	1

We set $q = p - k$ and generate $r = 100$ replicates of models M1-M7 in Table 1 with the specified distribution of $\mathbf{X}$ and sample size n. We estimate $\mathbf{B}$ using the four `ECVE` and three reference methods. The accuracy of the estimates is assessed using (6.2). The results are displayed in Tables 2-8. In M1, which is taken from [25], even though the mean subspace agrees with the central subspace,

i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$, mean subspace estimation methods, such as `mean MAVE` and `CVE`, fail because of the unboundedness of the link function $g(x) = 1/x$. In contrast, all four `ECVE` methods and `csMAVE` succeed in identifying the minimal dimension reduction subspace. On the other hand, both `reMAVE` procedures give unsatisfactory results, as can be seen in Table 2. Overall, `Fourier` is the best performing ensemble method.

`csMAVE` is the best performing method for the two dimensional mean subspace model M2 in Table 3. M3 is the same as model (6.1), where the mean subspace is a proper subset of the central subspace. In Table 4 we see that `Indicator_weighted` and `csMAVE` are the best performers and are roughly on par. In M4, the two dimensional mean subspace, which also determines the heteroskedasticity, agrees with the central subspace. Table 5 shows that this model is quite challenging for all methods, and only `Indicator_weighted` and `csMAVE` give satisfactory results, with `Indicator_weighted` the clear winner.

In M5, heteroskedasticity is induced by an interaction term and the three dimensional central subspace is a proper superset of the one dimensional mean subspace. M5 is quite challenging for all five methods, as can be seen In Table 6. When we increased the sample size to 800, the two weighted ensemble conditional variance estimators became the best performing methods followed by `csMAVE`.

M6 is a one dimensional pure central subspace model; i.e., the mean subspace is $\{0\}$. In Table 7, we see that for $n = 100$, the two weighted `ECVE`s are the best performing methods. For larger sample sizes, `csMAVE` is slightly more accurate than the `ECVE` methods.

M7 is the model where all ensemble conditional variance estimators clearly outperform all reference methods (Table 8). Here the one dimensional mean subspace coincides with the central subspace and the conditional first and second moments, $\mathbb{E}(Y^l \mid \mathbf{X})$ for $l = 1, 2$, are highly nonlinear periodic functions of the sufficient reduction.

7. Boston housing data

We apply the ensemble conditional variance estimator and `csMAVE` to the `Boston Housing` data set since the other reference methods are not competitive. This data set has been extensively used as a benchmark for assessing regression methods [see, for example, [16]], and is available in the `R`-package `mlbench`. The data contains 506 instances of 14 variables from the 1970 Boston census, 13 of which are continuous. The binary variable `chas`, indexing proximity to the Charles river, is omitted from the analysis since ensemble conditional variance estimation operates under the assumption of continuous predictors. The target variable is the median value of owner-occupied homes, `medv`, in $1,000. The 12 predictors are `crim` (per capita crime rate by town), `zn` (proportion of residential land zoned for lots over 25,000 sq.ft), `indus` (proportion of non-retail business acres per town), `nox` (nitric oxides concentration (parts per 10 million)), `rm` (average number of rooms per dwelling), `age` (proportion of owner-occupied units built

TABLE 2
Mean and standard deviation (in parenthesis) of estimation errors of M1

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	**0.172** (0.047)	0.201 (0.054)	0.248 (0.064)	0.265 (0.063)	0.210 (0.063)	0.843 (0.177)	0.972 (0.051)
I	200	**0.120** (0.029)	0.142 (0.037)	0.182 (0.045)	0.197 (0.049)	0.128 (0.037)	0.779 (0.266)	0.975 (0.055)
I	400	**0.079** (0.020)	0.091 (0.024)	0.126 (0.037)	0.136 (0.040)	0.080 (0.024)	0.678 (0.289)	0.975 (0.062)
II	100	**0.174** (0.038)	0.196 (0.049)	0.241 (0.055)	0.254 (0.056)	0.193 (0.059)	0.862 (0.164)	0.980 (0.038)
II	200	**0.110** (0.031)	0.127 (0.033)	0.170 (0.043)	0.182 (0.045)	0.121 (0.036)	0.781 (0.237)	0.974 (0.053)
II	400	**0.078** (0.021)	0.091 (0.026)	0.122 (0.031)	0.132 (0.033)	0.079 (0.020)	0.663 (0.305)	0.976 (0.059)
III	100	**0.187** (0.045)	0.218 (0.053)	0.256 (0.060)	0.263 (0.058)	0.204 (0.066)	0.822 (0.200)	0.975 (0.040)
III	200	**0.118** (0.031)	0.137 (0.038)	0.171 (0.043)	0.179 (0.042)	**0.118** (0.033)	0.685 (0.282)	0.967 (0.054)
III	400	0.082 (0.020)	0.101 (0.029)	0.127 (0.031)	0.132 (0.032)	**0.079** (0.022)	0.627 (0.306)	0.967 (0.065)

TABLE 3
Mean and standard deviation (in parenthesis) of estimation errors of M2

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.670 (0.089)	0.601 (0.135)	0.629 (0.130)	0.582 (0.140)	**0.575** (0.176)	0.699 (0.103)	0.683 (0.097)
I	200	0.478 (0.201)	0.388 (0.152)	0.436 (0.193)	0.407 (0.162)	**0.219** (0.136)	0.558 (0.178)	0.561 (0.179)
I	400	0.226 (0.153)	0.201 (0.074)	0.231 (0.127)	0.236 (0.111)	**0.098** (0.025)	0.285 (0.209)	0.288 (0.226)
II	100	0.663 (0.097)	0.652 (0.104)	0.687 (0.057)	0.658 (0.080)	**0.544** (0.176)	0.706 (0.108)	0.706 (0.063)
II	200	0.525 (0.171)	0.468 (0.171)	0.601 (0.127)	0.539 (0.148)	**0.182** (0.096)	0.576 (0.151)	0.659 (0.087)
II	400	0.267 (0.081)	0.307 (0.146)	0.375 (0.154)	0.357 (0.141)	**0.087** (0.021)	0.322 (0.203)	0.510 (0.215)
III	100	0.657 (0.104)	0.590 (0.148)	**0.530** (0.155)	0.542 (0.148)	0.603 (0.193)	0.742 (0.089)	0.708 (0.074)
III	200	0.421 (0.203)	0.367 (0.165)	0.306 (0.147)	0.336 (0.151)	**0.240** (0.193)	0.637 (0.144)	0.649 (0.130)
III	400	0.170 (0.110)	0.170 (0.071)	0.144 (0.053)	0.170 (0.063)	**0.089** (0.019)	0.453 (0.237)	0.482 (0.249)

prior to 1940), `dis` (weighted distances to five Boston employment centres), `rad` (index of accessibility to radial highways), `tax` (full-value property-tax rate per \$10,000), `ptratio` (pupil-teacher ratio by town), `lstat` (percentage of lower status of the population), and `b` stands for $1000(B - 0.63)^2$ where B is the proportion of blacks by town.

We analyze these data with the weighted and unweighted Fourier and indicator ensembles, and `csMAVE`. We compute unbiased error estimates by leave-one-out cross-validation. We estimate the sufficient reduction with the five methods from the standardized training set, estimate the forward model from the reduced training set using `mars`, multivariate adaptive regression splines [11], in the `R`-package `mda`, and predict the target variable on the test set. We report results for dimension $k = 1$. The analysis was repeated setting $k = 2$ with similar re-

TABLE 4
Mean and standard deviation (in parenthesis) of estimation errors of M3

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.744 (0.056)	0.657 (0.113)	0.668 (0.083)	**0.561** (0.142)	0.602 (0.147)	0.549 (0.142)	0.697 (0.084)
I	200	0.702 (0.061)	0.472 (0.177)	0.559 (0.147)	**0.369** (0.155)	0.374 (0.148)	**0.350** (0.102)	0.604 (0.086)
I	400	0.621 (0.148)	0.252 (0.102)	0.408 (0.177)	0.223 (0.064)	**0.203** (0.061)	0.221 (0.048)	0.547 (0.102)
II	100	0.751 (0.041)	0.698 (0.076)	0.683 (0.080)	**0.570** (0.136)	0.635 (0.136)	0.617 (0.139)	0.707 (0.085)
II	200	0.719 (0.040)	0.521 (0.163)	0.584 (0.111)	**0.355** (0.097)	0.387 (0.144)	0.376 (0.129)	0.631 (0.100)
II	400	0.686 (0.079)	0.267 (0.084)	0.452 (0.153)	0.252 (0.052)	**0.201** (0.045)	0.238 (0.072)	0.590 (0.099)
III	100	0.739 (0.073)	0.676 (0.106)	0.654 (0.105)	0.563 (0.150)	0.571 (0.120)	**0.501** (0.131)	0.666 (0.089)
III	200	0.704 (0.048)	0.546 (0.162)	0.523 (0.171)	0.368 (0.153)	0.330 (0.131)	**0.318** (0.083)	0.595 (0.091)
III	400	0.616 (0.151)	0.252 (0.113)	0.297 (0.106)	0.202 (0.055)	**0.179** (0.042)	0.215 (0.055)	0.551 (0.104)

TABLE 5
Mean and standard deviation (in parenthesis) of estimation errors of M4

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.836 (0.072)	0.794 (0.076)	0.774 (0.074)	**0.713** (0.105)	0.803 (0.087)	0.764 (0.091)	0.797 (0.092)
I	200	0.820 (0.066)	0.733 (0.094)	0.747 (0.060)	**0.545** (0.150)	0.685 (0.116)	0.631 (0.117)	0.758 (0.094)
I	400	0.782 (0.059)	0.633 (0.142)	0.710 (0.081)	**0.364** (0.129)	0.534 (0.155)	0.508 (0.142)	0.664 (0.117)
II	100	0.839 (0.067)	0.828 (0.064)	0.788 (0.062)	**0.751** (0.095)	0.818 (0.095)	0.789 (0.099)	0.805 (0.085)
II	200	0.834 (0.171)	0.781 (0.081)	0.759 (0.040)	0.660 (0.117)	0.701 (0.111)	**0.654** (0.116)	0.753 (0.093)
II	400	0.812 (0.059)	0.712 (0.097)	0.739 (0.038)	**0.511** (0.135)	0.544 (0.151)	0.521 (0.116)	0.681 (0.088)
III	100	0.838 (0.074)	0.815 (0.077)	0.764 (0.069)	**0.706** (0.108)	0.786 (0.109)	0.715 (0.116)	0.774 (0.086)
III	200	0.829 (0.071)	0.761 (0.099)	0.726 (0.083)	**0.544** (0.149)	0.676 (0.123)	0.616 (0.113)	0.735 (0.099)
III	400	0.796 (0.069)	0.646 (0.139)	0.669 (0.113)	**0.317** (0.110)	0.506 (0.146)	0.478 (0.122)	0.700 (0.085)

sults. Table 9 reports the first quantile, median, mean and third quantile of the out-of-sample prediction errors. The reductions estimated by the ensemble `CVE` methods achieve lower mean and median prediction errors than `csMAVE`. Also, both `ECVE` and `csMAVE` are approximately on par with the variable selection methods in [16, Section 8.3.3].

We plot the standardized response `medv` against the reduced `Fourier` and `csMAVE` predictors, $\mathbf{B}^T\mathbf{X}$, in Figure 4. The sufficient reductions are estimated using the entire data set. A particular feature of these data is that the response `medv` appears to be truncated as the highest median price of exactly \$50,000 is reported in 16 cases. Both methods pick up similar patterns, which is captured by the relatively high absolute correlation of the coefficients of the two reductions, $|\widehat{\mathbf{B}}_{\texttt{Fourier}}^T\widehat{\mathbf{B}}_{\texttt{csMAVE}}| = 0.786$. The coefficients of the reductions, $\widehat{\mathbf{B}}_{\texttt{Fourier}}$ and

TABLE 6
Mean and standard deviation (in parenthesis) of estimation errors of M5

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.705 (0.060)	**0.682** (0.067)	0.708 (0.060)	0.691 (0.056)	0.709 (0.069)	0.703 (0.067)	0.694 (0.061)
I	200	0.679 (0.061)	**0.634** (0.054)	0.688 (0.058)	0.642 (0.060)	0.687 (0.073)	0.691 (0.061)	0.693 (0.059)
I	400	0.644 (0.050)	**0.588** (0.047)	0.660 (0.056)	0.591 (0.061)	0.646 (0.082)	0.673 (0.064)	0.685 (0.060)
I	800	0.622 (0.032)	0.543 (0.078)	0.629 (0.035)	**0.493** (0.100)	0.553 (0.077)	0.608 (0.070)	0.668 (0.064)
II	100	0.712 (0.060)	**0.688** (0.069)	0.713 (0.051)	0.697 (0.057)	0.722 (0.054)	0.731 (0.052)	0.715 (0.054)
II	200	0.693 (0.058)	**0.669** (0.065)	0.694 (0.054)	**0.669** (0.057)	0.697 (0.064)	0.713 (0.054)	0.714 (0.060)
II	400	0.670 (0.054)	**0.614** (0.059)	0.681 (0.052)	0.633 (0.050)	0.687 (0.067)	0.700 (0.056)	0.705 (0.053)
II	800	0.660 (0.053)	**0.584** (0.045)	0.672 (0.052)	0.585 (0.055)	0.589 (0.074)	0.680 (0.067)	0.691 (0.053)
III	100	0.706 (0.062)	**0.687** (0.062)	0.703 (0.061)	0.691 (0.061)	0.724 (0.051)	0.731 (0.052)	0.720 (0.057)
III	200	0.701 (0.063)	**0.655** (0.069)	0.702 (0.058)	0.668 (0.074)	0.703 (0.080)	0.708 (0.067)	0.693 (0.060)
III	400	0.659 (0.062)	**0.603** (0.072)	0.664 (0.059)	0.604 (0.077)	0.682 (0.081)	0.684 (0.071)	0.701 (0.056)
III	800	0.657 (0.064)	0.562 (0.068)	0.651 (0.052)	**0.513** (0.109)	0.602 (0.087)	0.639 (0.076)	0.668 (0.057)

TABLE 7
Mean and standard deviation (in parenthesis) of estimation errors of M6

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.304 (0.092)	**0.294** (0.082)	0.492 (0.135)	0.299 (0.087)	0.539 (0.255)	0.577 (0.193)	0.888 (0.105)
I	200	0.217 (0.057)	0.213 (0.054)	0.329 (0.107)	0.205 (0.059)	**0.194** (0.061)	0.371 (0.094)	0.826 (0.125)
I	400	0.142 (0.036)	0.146 (0.035)	0.199 (0.069)	0.138 (0.039)	**0.114** (0.034)	0.259 (0.073)	0.796 (0.100)
II	100	0.308 (0.094)	**0.293** (0.073)	0.479 (0.129)	0.299 (0.086)	0.488 (0.248)	0.572 (0.193)	0.891 (0.115)
II	200	0.205 (0.058)	0.210 (0.057)	0.321 (0.095)	0.210 (0.058)	**0.192** (0.061)	0.343 (0.092)	0.842 (0.090)
II	400	0.144 (0.039)	0.150 (0.042)	0.190 (0.055)	0.142 (0.045)	**0.111** (0.032)	0.247 (0.081)	0.786 (0.114)
III	100	0.373 (0.152)	0.375 (0.175)	0.504 (0.143)	**0.322** (0.083)	0.562 (0.273)	0.546 (0.167)	0.889 (0.088)
III	200	0.226 (0.065)	0.230 (0.070)	0.340 (0.100)	**0.218** (0.060)	**0.218** (0.083)	0.370 (0.102)	0.838 (0.082)
III	400	0.149 (0.039)	0.151 (0.038)	0.194 (0.068)	0.146 (0.042)	**0.114** (0.032)	0.260 (0.088)	0.807 (0.111)

$\widehat{\mathbf{B}}_{\texttt{csMAVE}}$, are reported in Table 10. For the `Fourier` ensemble, the variables `rm` and `lstat` have the highest influence on the target variable `medv`. This agrees with the analysis in [16, Section 8.3.4] where it was found that these two variables are by far the most important using different variable selection techniques, such as random forests and boosted regression trees. In contrast, the reduction estimated by `csMAVE` has a lower coefficient for `rm` and higher ones for `crim` and `rad`.

TABLE 8
Mean and standard deviation (in parenthesis) of estimation errors of M7

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE	reMAVEf	reMAVEb
I	100	0.273 (0.169)	**0.237** (0.050)	0.241 (0.136)	0.252 (0.158)	0.790 (0.316)	0.968 (0.056)	0.973 (0.043)
I	200	0.160 (0.093)	0.159 (0.041)	**0.143** (0.083)	0.153 (0.093)	0.425 (0.391)	0.960 (0.115)	0.954 (0.120)
I	400	0.098 (0.024)	0.104 (0.025)	**0.088** (0.021)	0.102 (0.093)	0.127 (0.202)	0.770 (0.364)	0.742 (0.383)
II	100	**0.233** (0.057)	0.260 (0.134)	0.236 (0.142)	0.265 (0.185)	0.902 (0.219)	0.980 (0.0310)	0.975 (0.039)
II	200	0.154 (0.058)	0.176 (0.124)	**0.145** (0.093)	0.150 (0.094)	0.649 (0.414)	0.952 (0.125)	0.951 (0.138)
II	400	0.097 (0.025)	0.110 (0.094)	**0.087** (0.022)	0.099 (0.093)	0.295 (0.391)	0.837 (0.295)	0.868 (0.271)
III	100	0.274 (0.201)	0.303 (0.237)	**0.238** (0.160)	0.298 (0.242)	0.933 (0.163)	0.979 (0.037)	0.977 (0.041)
III	200	0.167 (0.120)	0.188 (0.159)	**0.159** (0.150)	0.167 (0.144)	0.678 (0.408)	0.971 (0.050)	0.973 (0.041)
III	400	0.100 (0.023)	0.116 (0.090)	**0.089** (0.023)	0.112 (0.129)	0.375 (0.431)	0.955 (0.130)	0.949 (0.146)

TABLE 9
Summary statistics of the out of sample prediction errors for the Boston Housing data obtained by LOO cross validation

	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
25% quantile	**0.766**	0.785	0.973	0.916	0.851
median	**3.323**	3.358	3.844	3.666	4.515
mean	19.971	19.948	19.716	**19.583**	24.309
75% quantile	11.129	10.660	11.099	**10.429**	16.521

TABLE 10
Rounded coefficients of the estimated reductions for $\widehat{\mathbf{B}}_{Fourier}$ and $\widehat{\mathbf{B}}_{csMAVE}$ from the full Boston Housing data

	crim	zn	indus	nox	rm	age	dis	rad	tax	ptratio	b	lstat
Fourier	0.21	-0.01	0.04	0.1	-0.62	0.16	0.2	0	0.2	0.27	-0.25	0.57
csMAVE	0.5	-0.05	-0.06	0.14	-0.27	0.11	0.24	-0.43	0.3	0.19	-0.15	0.51

8. Discussion

In this paper, we extend the *mean subspace* conditional variance estimation (CVE) of [10] to the ensemble conditional variance estimation (ECVE), which exhaustively estimates the *central subspace*, by applying the ensemble device introduced by [27]. We show that the new estimator is consistent for the central subspace. The regularity conditions for consistency require the joint distribution of the target variable and predictors, $(Y, \mathbf{X}^T)^T$, be sufficiently smooth. They are comparable to those under which the main competitor, csMAVE [25], is consistent.

ECVE identifies the central subspace via the orthogonal complement and thus circumvents the estimation and inversion of the variance matrix of the predictors $\mathbf{X}$. This renders the method formally applicable to settings where the sample size n is small, or smaller than p, the number of predictors, and leads to potential future research.

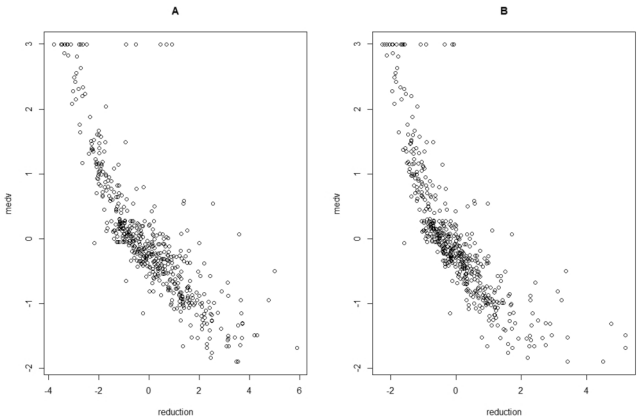

FIG 4. *Panel A: Y vs.* $\widehat{\mathbf{B}}_{Fourier}^T\mathbf{X}$. *Panel B: Y vs.* $\widehat{\mathbf{B}}_{csMAVE}^T\mathbf{X}$

`ECVE` is either on par with `csMAVE` or attains substantial performance improvement with respect to estimation accuracy in certain models. Yet, characterizing the defining features of the models for which the ensemble conditional variance estimation outperforms `csMAVE` entails further research. The nature of `ECVE`, i.e. the lack of closed-form solution and the lack of independence of all quantities in its calculation, presents many technical obstacles in deriving its statistical properties, such as the rate of convergence, estimation of the structural dimension, and the limiting distribution of the estimator. This can be seen from the technical difficulty of proving its consistency. An exception could be the dimension of the central subspace, $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$. We assumed k to be known throughout. Alternatively, the dimension can be estimated via cross-validation, as in [25] and [10], or information criteria.

Appendix A

For any $\mathbf{V} \in \mathcal{S}(p,q)$, defined in (2.1), we generically denote a basis of the orthogonal complement of its column space $\text{span}\{\mathbf{V}\}$, by $\mathbf{U}$. That is, $\mathbf{U} \in \mathcal{S}(p, p-q)$ such that $\text{span}\{\mathbf{V}\} \perp \text{span}\{\mathbf{U}\}$ and $\text{span}\{\mathbf{V}\} \cup \text{span}\{\mathbf{U}\} = \mathbb{R}^p$, $\mathbf{U}^T\mathbf{V} = \mathbf{0} \in \mathbb{R}^{(p-q)\times q}$, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{p-q}$. For any $\mathbf{x}, \mathbf{s}_0 \in \mathbb{R}^p$ we can always write

$$\mathbf{x} = \mathbf{s}_0 + \mathbf{P}_{\mathbf{V}}(\mathbf{x} - \mathbf{s}_0) + \mathbf{P}_{\mathbf{U}}(\mathbf{x} - \mathbf{s}_0) = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2 \tag{A.1}$$

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

The following theorem will be used in establishing the main result of this paper, which obtains the *exhaustive* sufficient reduction of the conditional distribution of Y given the predictor vector $\mathbf{X}$.

Theorem A.1. *Assume (E.1) and (E.2) hold, in particular model* (1.1) *holds. Let* $\widetilde{\mathbf{B}}$ *be a basis of* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$*; i.e.* $\text{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$.

Then, for any $f \in \mathcal{F}$ for which the statements in assumption (E.4) holds,

$$f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}, \tag{A.2}$$

with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and $g : \mathbb{R}^{k_t} \to \mathbb{R}$ is a twice continuously differentiable function, where $k_t = \dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})})$.

By Theorem A.1, any response Y can be written as an additive error via the decomposition (A.2). The predictors and the additive error term are only required to be conditionally uncorrelated in model (A.2). The *conditional variance estimator* [10] also estimated $\widetilde{\mathbf{B}}$ in (A.2) but under the more restrictive condition of predictor and error independence.

Proof of Theorem A.1.

$$\begin{aligned} f(Y) &= \mathbb{E}\left(f(Y) \mid \mathbf{X}\right) + \underbrace{f(Y) - \mathbb{E}\left(f(Y) \mid \mathbf{X}\right)}_{\tilde{\epsilon}} = \mathbb{E}\left(f(Y) \mid \mathbf{X}\right) + \tilde{\epsilon} \\ &= \mathbb{E}\left(f(Y)|\widetilde{\mathbf{B}}^T\mathbf{X}\right) + \tilde{\epsilon} = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon} \end{aligned}$$

where $g(\widetilde{\mathbf{B}}^T\mathbf{X}) = \mathbb{E}\left(f(Y)|\widetilde{\mathbf{B}}^T\mathbf{X}\right)$. By the tower property of the conditional expectation, $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = \mathbb{E}(f(Y) \mid \mathbf{X}) - \mathbb{E}(\mathbb{E}(f(Y) \mid \mathbf{X}) \mid \mathbf{X}) = \mathbb{E}(f(Y) \mid \mathbf{X}) - \mathbb{E}(f(Y) \mid \mathbf{X}) = \mathbf{0}$. The function g is twice continuous differentiable by (E.4). □

Theorem A.2. *Assume (E.1) and (E.2) hold. Let $\mathcal{F}$ be an ensemble, $\mathbf{s}_0 \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$, $\mathbf{V} \in S(p,q)$ defined in (2.1). Then, for any $f \in \mathcal{F}$ for which the statements in assumption (E.4) holds,*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1^2(\mathbf{V}, \mathbf{s}_0, f) + \mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) \tag{A.3}$$

where

$$\mu_l(\mathbf{V}, \mathbf{s}_0, f) = \int_{\mathbb{R}^q} g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1)^l \frac{f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} d\mathbf{r}_1 = \frac{t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)}, \tag{A.4}$$

for g given in (A.2) with

$$t^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = \int_{\mathbb{R}^q} g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1, \tag{A.5}$$

and

$$\begin{aligned} &\mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) \\ &= \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1 / \int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r} = \frac{\tilde{h}(\mathbf{V}, \mathbf{s}_0, f)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)} \end{aligned} \tag{A.6}$$

with $\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x})$ *and* $\tilde{h}(\mathbf{V}, \mathbf{s}_0, f) = \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1$. *Further assume* $h(\cdot)$ *to be continuous, then* $L^*(\mathbf{V}, f_t)$ *in* (3.3) *is well defined and continuous,*

$$\mathbf{V}_q^t = \operatorname{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L^*(\mathbf{V}, f_t) \tag{A.7}$$

is well defined, and the conditional variance estimator of the transformed response $f_t(Y)$ *identifies* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$,

$$\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \operatorname{span}\{\mathbf{V}_q^t\}^{\perp}. \tag{A.8}$$

Proof of Theorem A.2. The density of $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$ is given by

$$f_{\mathbf{X}|\mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}}(\mathbf{r}_1) = \frac{f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}} \tag{A.9}$$

where $\mathbf{X}$ is the p-dimensional continuous random covariate vector with density $f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{s}_0 \in \operatorname{supp}(f_{\mathbf{X}}) \subset \mathbb{R}^p$, and $\mathbf{V}$ belongs to the Stiefel manifold $\mathcal{S}(p, q)$ defined in (2.1). Equation (A.9) follows from Theorem 3.1 of [19] 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, which in turnguarantees the existence of the regular conditional probability of $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$ [see also [9]]. Further, the measure is concentrated on the affine subspace $\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\} \subset \mathbb{R}^p$ with density (A.9), which follows from Definition 8.38, Theorem 8.39 of [18], the orthogonal decomposition (A.1), and the continuity of $f_{\mathbf{X}}$ (E.2).

By assumption (E.1), $Y = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon)$ with $\epsilon \perp\!\!\!\perp \mathbf{X}$. Assume $f \in \mathcal{F}$ for which assumption (E.4) holds and let $\widetilde{\mathbf{B}}$ be a basis of $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$; that is, $\operatorname{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$. By Theorem A.1, $f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}$, with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and g twice continuously differentiable. Therefore,

$$\begin{aligned}
&\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\mathrm{ar}\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&= \mathbb{V}\mathrm{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) + 2\mathrm{cov}\left(\tilde{\epsilon}, g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&\qquad + \mathbb{V}\mathrm{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&= \mathbb{V}\mathrm{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) + \mathbb{V}\mathrm{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right)
\end{aligned} \tag{A.10}$$

The covariance term in (A.10) vanishes since

$$\begin{aligned}
&\mathrm{cov}\left(\tilde{\epsilon}, g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&= \mathbb{E}\left(\underbrace{\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X})}_{=0} g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&- E\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \mathbb{E}\left(\underbrace{\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X})}_{=0} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) = 0
\end{aligned}$$

i.e. the sigma field generated by $\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ is a subset of that generated by $\mathbf{X}$. By the same argument and using (A.9)

$$\begin{aligned}&\mathbb{V}\text{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \\ &= \mathbb{E}(\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \mathbb{E}(h(\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \\ &= \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) \times f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1 / t^{(0)}(\mathbf{V}, \mathbf{s}_0, f))\end{aligned}$$

where $\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x})$. Using (A.9) again for the first term in (A.10) obtains formula (A.3) and (A.6).

To see that (3.2), (A.3), and (A.6) are well defined and continuous, let $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) = g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})$ for $l = 1, 2$ or $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) = h(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})$ (for (A.6)) which are continuous by assumption. In consequence, the parameter depending integrals (A.5) and (A.6) are well defined and continuous if (1) $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \cdot)$ is integrable for all $\mathbf{V} \in \mathcal{S}(p, q), \mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$, (2) $\tilde{g}(\cdot, \cdot, \mathbf{r})$ is continuous for all $\mathbf{r}$, and (3) there exists an integrable dominating function of $\tilde{g}$ that does not depend on $\mathbf{V}$ and $\mathbf{s}_0$ [see [15, p. 101]].

Furthermore, for some compact set $\mathcal{K}$, $t^{(l)}(\mathbf{V}, \mathbf{s}_0) = \int_{\mathcal{K}} \tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) d\mathbf{r}$, since $\text{supp}(f_{\mathbf{X}})$ is compact by (E.2). The function $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r})$ is continuous in all inputs by the continuity of g (E.4) and $f_{\mathbf{X}}$ by (E.2), and therefore it attains a maximum. In consequence, all three conditions are satisfied so that $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$ is well defined and continuous. By the same argument (A.6) 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 interior points $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ by the continuity of $f_{\mathbf{X}}$, convexity of the support and $\mathbf{\Sigma}_{\mathbf{x}} > 0$. Then, $\tilde{L}(\mathbf{V}, \mathbf{s}_0, f)$ in (A.3) is continuous, which results in (3.3) also being well defined and continuous by virtue of it being a parameter depending integral following the same arguments as above. Moreover, (A.7) exists as the minimizer of a continuous function over the compact set $\mathcal{S}(p, q)$. Then, (3.3) can be writen as

$$\begin{aligned}L^*(\mathbf{V}, f) = &\mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}}\left(\mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2\right) \\ &+ \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}}\left(\mathbb{V}\text{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right)\right)\end{aligned} \tag{A.11}$$

where $\mathbf{s}_0 \sim \mathbf{X}$ signifies that $\mathbf{s}_0$ is distributed as $\mathbf{X}$ and the expectation is with respect to the distribution of $\mathbf{s}_0$.

It now suffices to show that the second term on the right hand side of (A.11) is constant with respect to $\mathbf{V}$. By the law of total variance,

$$\begin{aligned}\mathbb{V}\text{ar}(\tilde{\epsilon}) &= \mathbb{E}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) + \mathbb{V}\text{ar}\left(\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) \\ &= \mathbb{E}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right)\end{aligned} \tag{A.12}$$

since $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \mathbb{E}(\underbrace{\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X})}_{=0} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = 0$. Inserting (A.12) into (A.11) obtains

$$L^*(\mathbf{V}, f_t) = \mathbb{E}\left(\mu_2(\mathbf{V}, \mathbf{X}, f_t) - \mu_1(\mathbf{V}, \mathbf{X}, f_t)^2\right) + \mathbb{V}\text{ar}(\tilde{\epsilon})$$

$$= \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}} \left(\mathbb{V}\text{ar} \left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \right) \right) + \mathbb{V}\text{ar}(\tilde{\epsilon}) \tag{A.13}$$

Next we show that (3.3), or, equivalently (A.13)), attains its minimum at $\mathbf{V} \perp \widetilde{\mathbf{B}}$. Let $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ and $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q) \in \mathbb{R}^{p \times q}$, so that $\mathbf{v}_u \in \text{span}\{\mathbf{B}\}$ for some $u \in \{1, ..., q\}$. Since $\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \iff \mathbf{X} = \mathbf{s}_0 + \mathbf{P}_{\mathbf{V}}(\mathbf{X} - \mathbf{s}_0)$, by the first term in (A.13)

$$\begin{aligned} &\mathbb{V}\text{ar} \left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \right) \\ &= \mathbb{V}\text{ar} \left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \right) \\ &= \mathbb{V}\text{ar} \left(g(\widetilde{\mathbf{B}}^T \mathbf{s}_0 + \widetilde{\mathbf{B}}^T \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \right) \geq 0 \end{aligned} \tag{A.14}$$

If (A.14) is positive, i.e. $\widetilde{\mathbf{B}}^T \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \neq 0$ with positive probability, then the lower bound is not attained. If it is zero; i.e., for $\mathbf{V}$ such that $\mathbf{V}$ and $\widetilde{\mathbf{B}}^T$ are orthogonal, then $L^*(\mathbf{V}, f) = \mathbb{V}\text{ar}(\tilde{\epsilon})$. Since $\mathbf{s}_0$ is arbitrary yet constant, the same inequality holds for (3.3); that is, (3.3) attains its minimum for $\mathbf{V}$ such that $\mathbf{V}$ and $\widetilde{\mathbf{B}}^T$ are orthogonal. Since $\text{span}\{\widetilde{\mathbf{B}}^T\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$, (A.7) follows. □

[10] assumed model $Y = g(\mathbf{B}^T \mathbf{X}) + \epsilon$ with $\epsilon \perp\!\!\!\perp \mathbf{X}$, which implies $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$, and showed that the *conditional variance estimator (CVE)* can identify $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ at the population level.

Theorem A.2 extends this result to obtain that the *conditional variance estimator (CVE)* identifies the *mean subspace* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ also in models of the form $Y = g(\mathbf{B}^T \mathbf{X}) + \tilde{\epsilon}$, where $\tilde{\epsilon}$ is simply conditionally uncorrelated with $\mathbf{X}$. This allows CVE to apply to problems where the *mean subspace* is a proper subset of the *central subspace*, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subsetneq \mathcal{S}_{Y|\mathbf{X}}$.

$\mathbf{V}_q^t$ in (A.7) is not unique since for all orthogonal $\mathbf{O} \in \mathbb{R}^{q \times q}$, $L^*(\mathbf{V}_q^t \mathbf{O}, f_t) = L^*(\mathbf{V}_q^t, f_t)$ as $L^*(\mathbf{V}_q^t, f_t)$ depends on $\mathbf{V}_q^t$ only through $\text{span}\{\mathbf{V}_q^t\}$ by (3.1). Nevertheless, it is a unique minimizer over the Grassmann manifold $Gr(p, q)$ in (2.2). To see this, suppose $\mathbf{V} \in \mathcal{S}(p, q)$ is an arbitrary basis of a subspace $\mathbf{M} \in Gr(p, q)$. We can identify $\mathbf{M}$ through the projection $\mathbf{P}_{\mathbf{M}} = \mathbf{V}\mathbf{V}^T$. By (A.1), we write $\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$. Application of the Fubini-Tonelli Theorem yields

$$\begin{aligned} \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{M}}, \mathbf{s}_0, f) &= \int_{\text{supp}(f_{\mathbf{X}})} g(\mathbf{B}^T \mathbf{s}_0 + \mathbf{B}^T \mathbf{P}_{\mathbf{M}} \mathbf{x})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{P}_{\mathbf{M}} \mathbf{x}) d\mathbf{x} \\ &= t^{(l)}(\mathbf{V}, \mathbf{s}_0, f) \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} d\mathbf{r}_2. \end{aligned} \tag{A.15}$$

Therefore $\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{M}}, \mathbf{s}_0, f)/\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{M}}, \mathbf{s}_0, f) = t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)/t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)$ and $\mu_l(\cdot, \mathbf{s}_0, f)$ in (A.4) can also be viewed as a function from $Gr(p, q)$ to $\mathbb{R}$.

Proof of Theorem 3.1. Under assumptions (E.1), (E.2), and (E.3), (3.2) is well defined and continuous by arguments analogous to those in the proof of Theorem A.2. Therefore (3.4) exists as a minimizer of a continuous function over the compact set $\mathcal{S}(p, q)$.

To show $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{V}_q\}^\perp$, let $\tilde{\mathcal{S}} \neq \mathcal{S}_{Y|\mathbf{X}}$ with $\dim(\tilde{\mathcal{S}}) = \dim(\mathcal{S}_{Y|\mathbf{X}}) = k$. Also, let $\mathbf{Z} \in \mathbb{R}^{p\times(p-k)}$ be an orthonormal base of $\tilde{\mathcal{S}}^\perp$. Suppose $L_{\mathcal{F}}(\mathbf{Z}) = \min_{V\in\mathcal{S}(p,p-k)} L_{\mathcal{F}}(\mathbf{V})$. By (A.7) and (A.8) in Theorem A.2, $L^*(\mathbf{V}, f_t)$, considered as a function from $\mathbb{R}^{p\times(p-k_t)}$, is minimized by an orthonormal base of $\mathcal{S}^\perp_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ with $p - k_t$ elements, where $k_t = \dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}) \leq k$. By (E.1), $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$. As in the proof of Theorem A.2, we obtain that $L^*(\mathbf{V}, f_t)$, as a function from $\mathbb{R}^{p\times(p-k)}$, is minimized by an orthonormal base $\mathbf{U} \in \mathbb{R}^{p\times(p-k)}$ of $\text{span}\{\mathbf{B}\}^\perp$.

Since $\tilde{\mathcal{S}} = \text{span}\{\mathbf{Z}\} \neq \text{span}\{\mathbf{U}\} = \mathcal{S}_{Y|\mathbf{X}}$, we can rearrange the bases $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ such that $\text{span}\{\mathbf{U}_1\} = \text{span}\{\mathbf{Z}_1\}$ and $\text{span}\{\mathbf{U}_2\} \neq \text{span}\{\mathbf{Z}_2\}$. Since $\mathcal{F}$ characterizes $\mathcal{S}_{Y|\mathbf{X}}$, the set $A = \{t \in \Omega_T : \text{span}\{\mathbf{U}_2\} \subseteq \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}\}$ is non empty and by (E.3) A is not a null set with respect to the probability measure F_T.

Thus,

$$\begin{aligned}&\min_{V\in\mathcal{S}(p,p-k)} L_{\mathcal{F}}(\mathbf{V}) = L_{\mathcal{F}}(\mathbf{Z}) = \mathbb{E}_{t\sim F_T}\left(L^*(\mathbf{Z}, f_t)\right)\\ &= \int_A \underbrace{L^*(\mathbf{Z}, f_t)}_{>L^*(\mathbf{U},f_t)} dF_T(t) + \int_{A^c} \underbrace{L^*(\mathbf{Z}, f_t)}_{=L^*(\mathbf{U},f_t)} dF_T(t) > \mathbb{E}_{t\sim F_T}\left(L^*(\mathbf{U}, f_t)\right),\end{aligned}$$

which contradicts our assumption that $L_{\mathcal{F}}(\mathbf{Z}) = \min_{V\in\mathcal{S}(p,p-k)} L_{\mathcal{F}}(\mathbf{V})$. □

Appendix B

Next we introduce notation and auxiliary lemmas for the proof of Theorem 5.1. We suppose all assumptions of Theorem 5.1 hold. We generically use the letter "C" to denote constants.

Suppose f is an arbitrary element of $\mathcal{F}$ and let

$$\tilde{Y}_i = f(Y_i) = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i) + \tilde{\epsilon}_i \tag{B.1}$$

with $\text{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(\tilde{Y}|\mathbf{X})} = \mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}$. Condition (E.4) yields that g is twice continuously differentiable, and $\mathbb{E}(|\tilde{Y}|^8) < \infty$. Since f is fixed, we suppress it in $t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)$ and $\tilde{h}(\mathbf{V}, \mathbf{s}_0, f)$, so that

$$t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = \frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^n K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) \tilde{Y}_i^l, \tag{B.2}$$

which is the sample version of (A.5) for $l = 0, 1, 2$. Eqn. (4.3) can be expressed as

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)}, \tag{B.3}$$

Lemma 2. *Assume (E.2) and (K.1) hold. For a continuous function g, we let* $Z_n(\mathbf{V},\mathbf{s}_0) = \left(\sum_i g(\mathbf{X}_i)^l K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)\right)/(nh_n^{(p-q)/2})$. *Then,*

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

where $\tilde{g}(\mathbf{r}_1,\mathbf{r}_2) = g(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\mathbf{r}_2)$, $\mathbf{x} = \mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\mathbf{U}\mathbf{r}_2$ in (A.1).

Proof of Lemma 2. By (A.1), $\|\mathbf{P}_{\mathbf{U}}(\mathbf{x}-\mathbf{s}_0)\|^2 = \|\mathbf{U}\mathbf{r}_2\|^2 = \|\mathbf{r}_2\|^2$. Further

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

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

Lemma 3. *Assume (E.1), (E.2), (E.3), (E.4), (H.1) and (K.1) hold. Then, there exists a constant $C>0$, such that*

$$\mathbb{V}\mathrm{ar}\left(nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V},\mathbf{s}_0,f)\right) \le nh_n^{(p-q)/2} C$$

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

Proof of Lemma 3. Since a continuous function attains a finite maximum over a compact set, $\sup_{\mathbf{x}\in\mathrm{supp}(f_{\mathbf{X}})} |g(\widetilde{\mathbf{B}}^T\mathbf{x})| < \infty$. Therefore,

$$|\tilde{Y}_i| \le |g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)| + |\tilde{\epsilon}_i| \le \sup_{\mathbf{x}\in\mathrm{supp}(f_{\mathbf{X}})} |g(\widetilde{\mathbf{B}}^T\mathbf{x})| + |\tilde{\epsilon}_i| = C + |\tilde{\epsilon}_i|$$

and $|\tilde{Y}_i|^{2l} \le \sum_{u=0}^{2l} \binom{2l}{u} C^u |\tilde{\epsilon}_i|^{2l-u}$. Since $(\tilde{Y}_i, \mathbf{X}_i)$ are i.i.d.,

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}\left(nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V},\mathbf{s}_0,f)\right) = n\mathbb{V}\mathrm{ar}\left(\tilde{Y}^l K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\right) \\
&\le n\mathbb{E}\left(\tilde{Y}^{2l} K^2\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\right) = n\mathbb{E}\left(|\tilde{Y}|^{2l} K^2\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\right) \\
&\le n\sum_{u=0}^{2l}\binom{2l}{u} C^u \mathbb{E}\left(|\tilde{\epsilon}_i|^{2l-u} K^2\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\right)
\end{aligned}$$

$$= n \sum_{u=0}^{2l} \binom{2l}{u} C^u \mathbb{E}\left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) K^2 \left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n} \right) \right) \tag{B.4}$$

for $l = 0, 1, 2$. Let $\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) = g_{2l-u}(\mathbf{X}_i)$ for a continuous (by assumption) function $g_{2l-u}(\cdot)$ with finite moments for $l = 0, 1, 2$ by the compactness of $\text{supp}(f_{\mathbf{X}})$. Using Lemma 2 with

$$Z_n(\mathbf{V}, \mathbf{s}_0) = \frac{1}{n h_n^{(p-q)/2}} \sum_i g_{2l-u}(\mathbf{X}_i) K^2 \left(d_i(\mathbf{V}, \mathbf{s}_0)/h_n \right),$$

where $K^2(\cdot)$ fulfills (K.1), we calculate

$$\begin{aligned} \mathbb{E}\left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) K^2 \left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n} \right) \right) &= h_n^{(p-q)/2} \mathbb{E}(Z_n(\mathbf{V}, \mathbf{s}_0)) \\ &= h_n^{(p-q)/2} \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} K^2(\|\mathbf{r}_2\|^2) \times \\ \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} g_{2l-u}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \\ &\leq h_n^{(p-q)/2} C \end{aligned} \tag{B.5}$$

since all integrands in (B.5) are continuous and over compact sets by (E.2) and the continuity of $g_{2l-u}(\cdot)$ and $K(\cdot)$, so that the integral can be upper bounded by a finite constant C. Inserting (B.5) into (B.4) yields

$$\mathbb{V}\text{ar}\left(n h_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f) \right) \leq n h_n^{(p-q)/2} \underbrace{\sum_{u=0}^{2l} \binom{2l}{u} C^u C}_{=C} = n h_n^{(p-q)/2} C \tag{B.6}$$

□

In Lemma 4 we show that $d_i(\mathbf{V}, \mathbf{s}_0)$ in (4.1) is Lipschitz in its inputs under assumption (E.2).

Lemma 4. *Under assumption (E.2) there exists a constant $0 < C_2 < \infty$ such that for all $\delta > 0$ and $\mathbf{V}, \mathbf{V}_j \in \mathcal{S}(p, q)$ with $\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| < \delta$ and for all $\mathbf{s}_0, \mathbf{s}_j \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$ with $\|\mathbf{s}_0 - \mathbf{s}_j\| < \delta$,*

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

for $d_i(\mathbf{V}, \mathbf{s}_0)$ given by (4.1)

Proof of Lemma 4.

$$\begin{aligned} |d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq \left| \|\mathbf{X}_i - \mathbf{s}_0\|^2 - \|\mathbf{X}_i - \mathbf{s}_j\|^2 \right| + \\ \left| \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{P}_{\mathbf{V}}(\mathbf{X}_i - \mathbf{s}_0) \rangle - \langle \mathbf{X}_i - \mathbf{s}_j, \mathbf{P}_{\mathbf{V}_j}(\mathbf{X}_i - \mathbf{s}_j) \rangle \right| = I_1 + I_2 \end{aligned} \tag{B.7}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^p$. We bound the first term on the right hand side of (B.7) as follows using $\|\mathbf{X}_i\| \leq \sup_{z \in \text{supp}(f_{\mathbf{X}})} \|z\| = C_1 < \infty$ with probability 1 by (E.2).

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

by Cauchy-Schwartz and the reverse triangular inequality for which $\left| \|\mathbf{s}_0\|^2 - \|\mathbf{s}_j\|^2 \right| = |\|\mathbf{s}_0\| - \|\mathbf{s}_j\|| \, (\|\mathbf{s}_0\| + \|\mathbf{s}_j\|) \leq \|\mathbf{s}_0 - \mathbf{s}_j\| 2C_1$. The second term in (B.7) satisfies

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

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

To show Theorems 5.1 and 5, we use the **Bernstein inequality** [3]. Let $\{Z_i, i = 1, 2, \ldots\}$, be an independent sequence of bounded random variables with $|Z_i| \leq b$. Let $S_n = \sum_{i=1}^n Z_i$, $E_n = \mathbb{E}(S_n)$ and $V_n = \mathbb{V}\text{ar}(S_n)$. Then,

$$P(|S_n - E_n| > t) < 2 \exp\left(-\frac{t^2/2}{V_n + bt/3} \right) \tag{B.8}$$

Assumption (K.2) yields

$$|K(u) - K(u')| \leq K^*(u')\delta \tag{B.9}$$

for all u, u' with $|u - u'| < \delta \leq L_2$ and $K^*(\cdot)$ is a bounded and integrable kernel function [see [14]]. Specifically, if condition (1) of (K.2) holds, then $K^*(u) = L_1 1_{\{|u| \leq 2L_2\}}$. If condition (2) holds, then $K^*(u) = L_1 1_{\{|u| \leq 2L_2\}} + 1_{\{|u| > 2L_2\}} |u - L_2|^{-\nu}$.

Let $A = \mathcal{S}(p, q) \times \text{supp}(f_{\mathbf{X}})$. In Lemma 5 and 6 we show that (B.2) converges uniformly in probability to (A.5) by showing that the variance and bias terms vanish uniformly in probability, respectively.

Lemma 5. *Under the assumptions of Theorem 5.1,*

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) \right) \right| = O_P(a_n), \; l = 0, 1, 2 \tag{B.10}$$

Proof of Lemma 5. The proof proceeds in 3 steps: (i) truncation, (ii) discretization by covering $A = \mathcal{S}(p, q) \times \text{supp}(f_{\mathbf{X}})$, and (iii) application of Bernstein's inequality (B.8). If the function f in (B.1) is bounded, the truncation step and the assumption $a_n / h_n^{(p-q)/2} = O(1)$ are not needed.

(i) We let $\tau_n = a_n^{-1}$ and truncate $\tilde{Y}_i^l$ by τ_n as follows. We let

$$t_{n,\mathrm{trc}}^{(l)}(\mathbf{V},\mathbf{s}_0) = (1/nh_n^{(p-q)/2}) \sum_i K(\|\mathbf{P}_{\mathbf{U}}(\mathbf{X}_i - \mathbf{s}_0)\|^2/h_n)\tilde{Y}_i^l 1_{\{|\tilde{Y}_i|^l \le \tau_n\}} \quad \text{(B.11)}$$

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

$$\begin{aligned}\sup_{\mathbf{V}\times\mathbf{s}_0\in A} \left| t_n^{(l)}(\mathbf{V},\mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)\right| &\le M_1(\tilde{R}_n^{(l)} + \mathbb{E}\tilde{R}_n^{(l)}) \\ &+ \sup_{\mathbf{V}\times\mathbf{s}_0\in A} \left| t_{n,\mathrm{trc}}^{(l)}(\mathbf{V},\mathbf{s}_0) - \mathbb{E}\left(t_{n,\mathrm{trc}}^{(l)}(\mathbf{V},\mathbf{s}_0)\right)\right| \end{aligned} \quad \text{(B.12)}$$

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

$$\begin{aligned}\mathbb{E}\tilde{R}_n^{(l)} &= \frac{1}{h_n^{(p-q)/2}}\mathbb{E}\left(|\tilde{Y}_i|^l 1_{\{|\tilde{Y}_i|^l > \tau_n\}}\right) \le \frac{1}{h_n^{(p-q)/2}}\sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})}\sqrt{\mathbb{P}(|\tilde{Y}_i|^l > \tau_n)} \\ &\le \frac{1}{h_n^{(p-q)/2}}\sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})}\left(\frac{\mathbb{E}(|\tilde{Y}_i|^{4l})}{a_n^{-4}}\right)^{1/2} = o(a_n), \end{aligned} \quad \text{(B.13)}$$

where the last equality uses the assumption $a_n/h_n^{(p-q)/2} = O(1)$ and the expectations are finite due to (E.4) for $l = 0, 1, 2$. No truncation is needed for $l = 0$ or if $\tilde{Y}_i = f(Y_i) \le \sup_{f\in\mathcal{F}} |f(Y_i)| < C < \infty$.

Therefore, the first two terms of the right hand side of (B.12) converge to 0 with rate a_n by (B.13) and Markov's inequality. From this point on, $\tilde{Y}_i$ will denote the truncated version $\tilde{Y}_i 1_{\{|\tilde{Y}_i| \le \tau_n\}}$ and we do not distinguish the truncated from the untruncated $t_n(\mathbf{V},\mathbf{s}_0)$ since this truncation results in an error of magnitude a_n.

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

$$|d_i(\mathbf{V},\mathbf{s}_0) - d_i(\mathbf{V}_j,\mathbf{s}_j)| \le C_2\delta_n \quad \text{(B.14)}$$

for d_i in (4.1). Under (K.2), which implies (B.9), inequality (B.14) yields

$$\left| K\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right) - K\left(\frac{d_i(\mathbf{V}_j,\mathbf{s}_j)}{h_n}\right)\right| \le K^*\left(\frac{d_i(\mathbf{V}_j,\mathbf{s}_j)}{h_n}\right) C_2 a_n \quad \text{(B.15)}$$

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

Define $r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) = (1/nh_n^{(p-q)/2}) \sum_{i=1}^n K^*(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)|\tilde{Y}_i|^l$. For notational convenience we next drop the dependence on l and j and observe that (B.15) yields

$$|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2 a_n r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) \tag{B.16}$$

Since K^* fulfills (K.1) except for continuity, an analogous argument as in the proof of Lemma 2 yields that $\mathbb{E}\left(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right) < \infty$. By subtracting and adding $t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)$, $\mathbb{E}(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$, the triangular inequality, (B.16) and integrability of r_n^l, we obtain

$$\begin{aligned}
&\left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \leq \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right| \\
&+ \left|\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq C_2 a_n \left(|r_n| + |\mathbb{E}\left(r_n\right)|\right) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq C_2 a_n (|r_n - \mathbb{E}(r_n)| + 2|\mathbb{E}(r_n)|) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq 2C_3 a_n + |r_n - \mathbb{E}(r_n)| + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right|
\end{aligned} \tag{B.17}$$

for any constant $C_3 > C_2\mathbb{E}(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$ and n such that $C_2 a_n \leq 1$, since $a_n^2 = o(1)$, which in turn yields that there exists $0 < C_3 < \infty$ such that (B.17) holds.

Since $\sup_{x\in A} f(x) = \max_{1\leq j\leq N} \sup_{x\in A_j} f(x) \leq \sum_{j=1}^N \sup_{x\in A_j} f(x)$ for any cover of A and continuous function f,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{V}\times\mathbf{s}_0\in A} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n\right) \\
&\leq \sum_{j=1}^N \mathbb{P}\left(\sup_{\mathbf{V}\times\mathbf{s}_0\in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n\right) \\
&\leq N \max_{1\leq j\leq N} \mathbb{P}\left(\sup_{\mathbf{V}\times\mathbf{s}_0\in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n\right) \qquad \text{(B.18)} \\
&\leq N\Bigg(\max_{1\leq j\leq N} \mathbb{P}\left(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n\right) \qquad \text{(B.19)} \\
&\qquad + \max_{1\leq j\leq N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\Bigg) \leq \\
&C\,\delta^{-(d+p)}\Bigg(\max_{1\leq j\leq N} \mathbb{P}\left(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n\right) \qquad \text{(B.20)} \\
&\qquad + \max_{1\leq j\leq N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\Bigg)
\end{aligned}$$

by the subadditivity of probability for the first inequality and (B.17) for the

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

Finally, we bound the first and second term in the last line of (B.18) by the Bernstein inequality (B.8). For the first term in the last line of (B.18), let $Z_i = Y_i^l K(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)$ and $S_n = \sum_i Z_i = nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)$. Then, Z_i are independent with $|Z_i| \leq b = M_1\tau_n = M_1/a_n$ by (K.1) and the truncation step (i). For $V_n = \mathbb{V}\mathrm{ar}(S_n)$, Lemma 3 yields $nh_n^{(p-q)/2}C \geq V_n$ with $C > 0$, and set $t = C_3 a_n nh_n^{(p-q)/2}$. The Bernstein inequality (B.8) yields

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

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

For the second term in the last line of (B.18), set $Z_i = Y_i^l K^*(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)$ in (B.8) and proceed similarly to obtain

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

By (H.1), $h_n^{(p-q)/4} \leq 1$ for n large and (H.2) implies $1/(nh_n^{(p-q)/2}) \leq 1$ for n large, therefore $h_n^{-1} \leq n^{2/(p-q)} \leq n^2$ since $p - q \geq 1$. Then, $\delta_n^{-1} = (a_n h_n)^{-1} \leq n^{1/2}h_n^{-1}h_n^{(p-q)/4} \leq n^{5/2}$. Therefore, (B.18) is smaller than $4\,C\,\delta_n^{-(d+p)}n^{-\gamma(C_3)} \leq 4Cn^{5(d+p)/2-\gamma(C_3)}$. For C_3 large enough, we have $5(d+p)/2 - \gamma(C_3) < 0$ and $n^{5(d+p)/2-\gamma(C_3)} \to 0$. This completes the proof. □

If we assume $|\tilde{Y}_i| < M_2 < \infty$ 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 Lemma 5 is no longer necessary.

Lemma 6. *Under (E.1), (E.2), (E.3), (E.4), (H.1), (K.1), and* $\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)\,d\mathbf{r}_2 = 1$,

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|t^{(l)}(\mathbf{V}, \mathbf{s}_0) + 1_{\{l=2\}}\tilde{h}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| = O(h_n), \quad l = 0, 1, 2 \tag{B.21}$$

where $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$ *and* $\tilde{h}(\mathbf{V}, \mathbf{s}_0)$ *are defined in Theorem A.2.*

Proof of Lemma 6. Let $\tilde{g}(\mathbf{r}_1, \mathbf{r}_2) = g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1 + \widetilde{\mathbf{B}}^T\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)$, where $\mathbf{r}_1, \mathbf{r}_2$ satisfy the orthogonal decomposition (A.1).

$$\mathbb{E}\left(t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)\right) = \mathbb{E}\left(K(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)\right)/h_n^{(p-q)/2}$$

$$\mathbb{E}(t_n^{(1)}(\mathbf{V},\mathbf{s}_0)) = \mathbb{E}\left(K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)\right)/h_n^{(p-q)/2} + \mathbb{E}\left(K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)\underbrace{\mathbb{E}(\tilde{\epsilon}_i \mid \mathbf{X})}_{=0}\right)/h_n^{(p-q)/2}$$

$$\mathbb{E}(t_n^{(2)}(\mathbf{V},\mathbf{s}_0)) = \mathbb{E}\left(K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)^2\right)/h_n^{(p-q)/2} + 2\mathbb{E}\left(K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)\underbrace{\mathbb{E}(\tilde{\epsilon}_i \mid \mathbf{X})}_{=0}\right)/h_n^{(p-q)/2} + \mathbb{E}\left(K(d_i(\mathbf{V},\mathbf{s}_0)/h_n)\underbrace{\mathbb{E}(\tilde{\epsilon}_i^2 \mid \mathbf{X})}_{=h(\mathbf{X}_i)}\right)/h_n^{(p-q)/2}$$

Then

$$\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right) = \int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\tilde{g}(\mathbf{r}_1, {h_n}^{1/2}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \tag{B.22}$$

holds by Lemma 2 for $l = 0, 1$. For $l = 2$, $\tilde{Y}_i^2 = g_i^2 + 2g_i\epsilon_i + \epsilon_i^2$ with $g_i = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)$ and can be handled as in the case of $l = 0, 1$. Plugging in (B.22) the second order Taylor expansion for some ξ in the neighborhood of 0, $\tilde{g}(\mathbf{r}_1, {h_n}^{1/2}\mathbf{r}_2) = \tilde{g}(\mathbf{r}_1, 0) + {h_n}^{1/2}\nabla_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1, 0)^T\mathbf{r}_2 + h_n\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1, \xi)\mathbf{r}_2$, yields

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

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

Lemma 7 follows directly from Lemmas 5 and 6 and the triangle inequality.

Lemma 7. *Suppose (E.1), (E.2), (E.3), (E.4), (K.1), (K.2), (H.1) hold. If* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$, *and* $a_n/h_n^{(p-q)/2} = O(1)$, *then for* $l = 0, 1, 2$

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

Theorem B.1. *Suppose (E.1), (E.2), (E.3), (E.4), (K.1), (K.2), (H.1) hold. Let* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$, $a_n/h_n^{(p-q)/2} = O(1)$, *then*

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|\bar{y}_l(\mathbf{V},\mathbf{s}_0) - \mu_l(\mathbf{V},\mathbf{s}_0) - 1_{\{l=2\}}\tilde{h}(\mathbf{V},\mathbf{s}_0)\right| = o_P(1), \quad l = 0,1,2$$

and

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|\tilde{L}_n(\mathbf{V},\mathbf{s}_0) - \tilde{L}(\mathbf{V},\mathbf{s}_0)\right| = o_P(1) \tag{B.23}$$

where $\bar{y}_l(\mathbf{V},\mathbf{s}_0)$, $\mu_l(\mathbf{V},\mathbf{s}_0)$, $\tilde{L}_n(\mathbf{V},\mathbf{s}_0)$ *and* $\tilde{L}(\mathbf{V},\mathbf{s}_0)$ *are defined in* (4.3), (A.4), (4.4) *and* (A.3), *respectively.*

Proof of Theorem B.1. Let $\delta_n = \inf_{\mathbf{V}\times\mathbf{s}_0\in A_n} t^{(0)}(\mathbf{V},\mathbf{s}_0)$, where $t^{(0)}(\mathbf{V},\mathbf{s}_0)$ is defined in (A.5), and $A_n = \mathcal{S}(p,q)\times\{\mathbf{x}\in\text{supp}(f_{\mathbf{X}}) : |\mathbf{x} - \partial\text{supp}(f_{\mathbf{X}})| \geq b_n\}$, where ∂C denotes the boundary of the set C and $|\mathbf{x} - C| = \inf_{\mathbf{r}\in C}|\mathbf{x}-\mathbf{r}|$, for a sequence $b_n \to 0$ so that $\delta_n^{-1}(a_n + h_n) \to 0$ for any bandwidth h_n that satisfies the assumptions. Then,

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

We consider the numerator and enumerator of (B.24) separately. By Lemma 7

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

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

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

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

Substituting in (B.24), we obtain

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

For $l=2$, $\tilde{Y}_i^2 = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)^2 + 2g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)\tilde{\epsilon}_i + \tilde{\epsilon}_i^2$, and (B.23) follows from (A.3). □

Lemma 8. *Under (E.1), (E.2), (E.4), there exists* $0 < C_5 < \infty$ *such that*

$$|\mu_l(\mathbf{V},\mathbf{s}_0) - \mu_l(\mathbf{V}_j,\mathbf{s}_0)| \leq C_5\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{B.25}$$

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

Proof. From the representation $\tilde{t}^{(l)}(\mathbf{P_V}, \mathbf{s}_0)$ in (A.15) instead of $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$, we consider $\mu_l(\mathbf{V}, \mathbf{s}_0) = \mu_l(\mathbf{P_V}, \mathbf{s}_0)$ as a function on the Grassmann manifold since $\mathbf{P_V} \in Gr(p, q)$. Then,

$$\begin{aligned}\left|\mu_l(\mathbf{P_V}, \mathbf{s}_0) - \mu_l(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)\right| &= \left|\frac{\tilde{t}^{(l)}(\mathbf{P_V}, \mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0)} - \frac{\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)}\right| \\ &\leq \frac{\sup|\tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0)|}{(\inf \tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0))^2}\left|\tilde{t}^{(l)}(\mathbf{P_V}, \mathbf{s}_0) - \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)\right| \\ &\quad + \frac{\sup \tilde{t}^{(l)}(\mathbf{P_V}, \mathbf{s}_0)}{(\inf \tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0))^2}\left|\tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0) - \tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)\right| \end{aligned} \tag{B.26}$$

with $\sup_{\mathbf{P_V} \in Gr(p,q)} \tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0) \in (0, \infty)$ and $\inf_{\mathbf{P_V} \in Gr(p,q)} \tilde{t}^{(0)}(\mathbf{P_V}, \mathbf{s}_0) \in (0, \infty)$ since $\tilde{t}^{(l)}$ is continuous, $\mathbf{\Sigma_X} > 0$ and $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ an interior point.

By (E.2) and (E.4), $\tilde{g}(\mathbf{x}) = g(\widetilde{\mathbf{B}}^T\mathbf{x})f_{\mathbf{X}}(\mathbf{x})$ is twice continuous differentiable and therefore Lipschitz continuous on compact sets. We denote its Lipschitz constant by $L < \infty$. Therefore,

$$\begin{aligned}\left|\tilde{t}^{(l)}(\mathbf{P_V}, \mathbf{s}_0) - \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)\right| &\leq \int_{\text{supp}(f_{\mathbf{X}})} \left|\tilde{g}(\mathbf{s}_0 + \mathbf{P_V}\mathbf{r}) - \tilde{g}(\mathbf{s}_0 + \mathbf{P}_{\mathbf{V}_j}\mathbf{r})\right| d\mathbf{r} \\ \leq L \int_{\text{supp}(f_{\mathbf{X}})} \|(\mathbf{P_V} - \mathbf{P}_{\mathbf{V}_j})\mathbf{r}\| d\mathbf{r} &\leq L\left(\int_{\text{supp}(f_{\mathbf{X}})} \|\mathbf{r}\| dr\right)\|\mathbf{P_V} - \mathbf{P}_{\mathbf{V}_j}\| \end{aligned} \tag{B.27}$$

where the last inequality is due to the sub-multiplicativity of the Frobenius norm and the integral being finite by (E.2). Plugging (B.27) in (B.26) and collecting all constants into C_5 yields (B.25). □

Proof of Theorem 5.1. By (4.5) and (3.2),

$$\begin{aligned}|L_n^*(\mathbf{V}, f) - L^*(\mathbf{V}, f)| \leq &\left|\frac{1}{n}\sum_i \left(\tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}(\mathbf{V}, \mathbf{X}_i, f)\right)\right| \\ &+ \left|\frac{1}{n}\sum_i \left(\tilde{L}(\mathbf{V}, \mathbf{X}_i, f) - \mathbb{E}(\tilde{L}(\mathbf{V}, \mathbf{X}, f))\right)\right| \end{aligned} \tag{B.28}$$

By Theorem B.1,

$$\begin{aligned}\left|\frac{1}{n}\sum_i \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}(\mathbf{V}, \mathbf{X}_i, f)\right| &\leq \sup_{\mathbf{V} \times \mathbf{s}_0 \in A}\left|\tilde{L}_n(\mathbf{V}, \mathbf{s}_0, f) - \tilde{L}(\mathbf{V}, \mathbf{s}_0, f)\right| \\ &= o_P(1)\end{aligned}$$

The second term in (B.28) converges to 0 almost surely for all $\mathbf{V} \in \mathcal{S}(p, q)$ by the strong law of large numbers. In order to show uniform convergence the same technique as in the proof of Theorem 5 is used. Let $B_j = \{\mathbf{V} \in \mathcal{S}(p, q) :$

$\|\mathbf{V}\mathbf{V}^T - \mathbf{V}_j\mathbf{V}_j^T\| \leq \tilde{a}_n\}$ be a cover of $\mathcal{S}(p,q) \subset \bigcup_{j=1}^N B_j$ with $N \leq C\,\tilde{a}_n^{-d} = C\,(n/\log(n))^{d/2} \leq C\,n^{d/2}$, where $d = \dim(\mathcal{S}(p,q))$ is defined in the proof of Theorem 5. By Lemma 8,

$$|\mu_l(\mathbf{V}, \mathbf{X}_i) - \mu_l(\mathbf{V}_j, \mathbf{X}_i)| \leq C_5 \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{B.29}$$

Let $G_n(\mathbf{V}, f) = \sum_i \tilde{L}(\mathbf{V}, \mathbf{X}_i, f)/n$ with $\mathbb{E}(G_n(V)) = L^*(\mathbf{V}, f)$. Using (B.29) and following the same steps as in the proof of Lemma 5 we obtain

$$\begin{aligned} |G_n(\mathbf{V}, f) - L^*(\mathbf{V}, f)| &\leq |G_n(\mathbf{V}, f) - G_n(\mathbf{V}_j, f)| \\ &\quad + |G_n(\mathbf{V}_j, f) - L^*(\mathbf{V}_j, f)| + |L^*(\mathbf{V}, f) - L^*(\mathbf{V}_j, f)| \\ &\leq 2C_6\tilde{a}_n + |G_n(\mathbf{V}_j, f) - L^*(\mathbf{V}_j, f)| \end{aligned} \tag{B.30}$$

for $\mathbf{V} \in B_j$ and some $C_6 > C_5$. Inequality (B.30) leads to

$$\begin{aligned} &\mathbb{P}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |G_n(\mathbf{V}, f) - L^*(\mathbf{V}, f)| > 3C_6\tilde{a}_n\right) \\ &\leq C\,N\,\mathbb{P}(\sup_{\mathbf{V}\in B_j} |G_n(\mathbf{V}, f) - L^*(\mathbf{V}, f)| > 3C_6\tilde{a}_n) \\ &\leq C\,n^{d/2}\mathbb{P}(|G_n(\mathbf{V}_j, f) - L^*(\mathbf{V}_j, f)| > C_6\tilde{a}_n) \\ &\leq C\,n^{d/2}n^{-\gamma(C_6)} \to 0 \end{aligned} \tag{B.31}$$

where the last inequality in (B.31) is due to (B.8) with $Z_i = \tilde{L}(\mathbf{V}_j, \mathbf{X}_i, f)$, which is bounded since $(\cdot, \cdot, f)$ is continuous on the compact set A, and $\gamma(C_6)$ a monotone increasing function of C_6 that can be made arbitrarily large by choosing C_6 accordingly. Therefore, $\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f) - L^*(\mathbf{V}, f)| \leq o_P(1) + O_P(\tilde{a}_n)$ which implies Theorem 5.1. □

Proof of Theorem 5.2. We apply [2, Thm 4.1.1] to obtain consistency of the conditional variance estimator. This theorem requires three conditions that guarantee the convergence of the minimizer of a sequence of random functions $L_n^*(\mathbf{P}_{\mathbf{V}}, f_t)$ to the minimizer of the limiting function $L^*(\mathbf{P}_{\mathbf{V}}, f_t)$; i.e., $\mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}_{k_t}^t\}^\perp} = \operatorname{argmin} L_n^*(\mathbf{P}_{\mathbf{V}}, f) \to \mathbf{P}_{\text{span}\{\mathbf{B}\}^\perp} = \operatorname{argmin} L^*(\mathbf{P}_{\mathbf{V}}, f_t)$ in probability. To apply the theorem three conditions have to be met: (1) The parameter space is compact; (2) $L_n^*(\mathbf{P}_{\mathbf{V}}, f_t)$ is continuous in $\mathbf{P}_{\mathbf{V}}$ and a measurable function of the data $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$, and (3) $L_n^*(\mathbf{P}_{\mathbf{V}}, f_t)$ converges uniformly to $L^*(\mathbf{P}_{\mathbf{V}}, f_t)$ and $L^*(\mathbf{P}_{\mathbf{V}}, f_t)$ attains a unique global minimum at $\mathcal{S}^\perp_{\mathbb{E}(f_t(Y)|\mathbf{X})}$.

Since $L_n^*(\mathbf{V}, f_t)$ depends on $\mathbf{V}$ only through $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^T$, $L_n^*(\mathbf{V}, f_t)$ can be considered as functions on the Grassmann manifold, which is compact, and the same holds true for $L^*(\mathbf{V}, f_t)$ by (A.15). Further, $L_n^*(\mathbf{V}, f_t)$ is by definition a measurable function of the data and continuous in $\mathbf{V}$ if a continuous kernel, such as the Gaussian, is used. Theorem 5.1 obtains the uniform convergence and Theorem A.2 that the minimizer is unique when $L(\mathbf{V})$ is minimized over the Grassmann manifold $G(p,q)$, since $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \text{span}\{\widetilde{\mathbf{B}}\}$ is uniquely identifiable and so is $\text{span}\{\widetilde{\mathbf{B}}\}^\perp$ (i.e. $\|\mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}_{k_t}^t\}} - \mathbf{P}_{\text{span}\{\widetilde{\mathbf{B}}\}}\| = \|\widehat{\mathbf{B}}_{k_t}^t(\widehat{\mathbf{B}}_{k_t}^t)^T - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T\| =$

$\|(\mathbf{I}_p - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T) - (\mathbf{I}_p - \widehat{\mathbf{B}}^t_{k_t}(\widehat{\mathbf{B}}^t_{k_t})^T)\| = \|\mathbf{P}_{\text{span}\{\widetilde{\mathbf{B}}\}^\perp} - \mathbf{P}_{\text{span}\{\widehat{\mathbf{B}}^t_{k_t}\}^\perp}\|$). Thus, all three conditions are met and the result is obtained. □

Proof of Theorem 5.3. Let $(t_j)_{j=1,\dots,m_n}$ be an i.i.d. sample from F_T and write

$$|L_{n,\mathcal{F}}(\mathbf{V}) - L_{\mathcal{F}}(\mathbf{V})| = \left|\frac{1}{m_n}\sum_{j=1}^{m_n}\left(L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})\right)\right| + \left|\frac{1}{m_n}\sum_{j=1}^{m_n}\left(L^*(\mathbf{V}, f_{t_j}) - \mathbb{E}_{t\sim F_T}(L^*(\mathbf{V}, f_t)\right)\right| \quad \text{(B.32)}$$

Then, $\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f_t) - L^*(\mathbf{V}, f_t)| \leq 8M^2$, by the triangle inequality and the assumption that $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$. That is, $L_n^*(\mathbf{V}, f_t)$ estimates a variance of a bounded response $f_t(Y) \in [-M, M]$ and is therefore bounded by the squared range $4M^2$ of $f_t(Y)$. The same holds true for $L^*(\mathbf{V}, f_t)$. Further, $8M^2$ is an integrable dominant function so that Fatou's Lemma applies.

Consider the first term on the right hand side of (B.32) and let $\delta > 0$. By Markov's and triangle inequalities and Fatou's Lemma,

$$\begin{aligned}
&\limsup_n \mathbb{P}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)}\left|\frac{1}{m_n}\sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})\right| > \delta\right) \\
&\leq \frac{1}{\delta}\limsup_n \mathbb{E}_{F_T}\left(\mathbb{E}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)}\left|\frac{1}{m_n}\sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})\right|\right)\right) \\
&\leq \frac{1}{\delta}\limsup_n \mathbb{E}_{F_T}\left(\frac{1}{m_n}\sum_{j=1}^{m_n}\mathbb{E}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)}|L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})|\right)\right) \\
&= \frac{1}{\delta}\limsup_n \mathbb{E}_{F_T}\left(\mathbb{E}\left(\sup_{\mathbf{V}\in\mathcal{S}(p,q)}|L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})|\right)\right) \\
&\leq \frac{1}{\delta}\mathbb{E}_{F_T}\left(\mathbb{E}\left(\limsup_n \sup_{\mathbf{V}\in\mathcal{S}(p,q)}|L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})|\right)\right) = \frac{1}{\delta}\mathbb{E}_{t\sim F_T}\left(\mathbb{E}(0)\right) = 0
\end{aligned}$$

since by Theorem 5.1 it holds $\limsup_n \sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})| = 0$. The first inequality results from applying the Markov inequality.

For the second term on the right hand side of (B.32) we apply Theorem 2 of [17] in [23, p. 40]:

Theorem B.2. *Let t_j be an i.i.d. sample and $L^*(\mathbf{V}, f_t) : \Theta \times \Omega_T \to \mathbb{R}$ where Θ is a compact subset of an euclidean space. $L^*(\mathbf{V}, f_t)$ is continuous in $\mathbf{V}$ and measurable in t by Theorem A.2. If $L^*(\mathbf{V}, f_{t_j}) \leq h(t_j)$, where $h(t_j)$ is integrable with respect to F_T, then*

$$\frac{1}{m_n}\sum_{j=1}^{m_n} L^*(\mathbf{V}, f_{t_j}) \longrightarrow \mathbb{E}_{F_T}\left(L^*(\mathbf{V}, f_t)\right) \text{ uniformly over } \mathbf{V} \in \Theta \text{ a.s. as } n \to \infty$$

Here $\mathbf{V} \in \mathcal{S}(p,q) = \Theta \subseteq \mathbb{R}^{pq}$, by $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$ and an analogous argument as for the first term in (B.32), $Z_j(\mathbf{V}) = L^*(\mathbf{V}, f_{t_j}) < 4M^2$. Therefore, $\mathbb{E}(\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |Z_j(\mathbf{V})|) < 4M^2$, which is integrable. Further, since t_j are an i.i.d. sample from F_T, $Z_j(\mathbf{V})$ is a i.i.d. sequence of random variables, $Z_j(\mathbf{V})$ is continuous in $\mathbf{V}$ by Theorem A.2 and the parameter space $\mathcal{S}(p,q)$ is compact. Then by Theorem B.2,

$$\sup_{\mathbf{V}\in\mathcal{S}(p,q)} \left| \frac{1}{m_n} \sum_{j=1}^{m_n} L^*(\mathbf{V}, f_{t_j}) - \mathbb{E}_{t\sim F_T}(L^*(\mathbf{V}, f_t)) \right| \longrightarrow 0 \quad \text{a.s. as } n \to \infty$$

if $\lim_{n\to\infty} m_n = \infty$. Putting everything together it follows that $\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_{n,\mathcal{F}}(\mathbf{V}) - L_{\mathcal{F}}(\mathbf{V})| \to 0$ in probability as $n \to \infty$. □

Proof of Theorem 5.4. The proof is directly analogous to the proof of Theorem 5.2. The uniform convergence of the target function $L_{n,\mathcal{F}}(\mathbf{V})$ is obtained by Theorem 5.3. The minimizer over $Gr(p,q)$ and its uniqueness derive from Theorem 3.1. □

Proof of Theorem 4.1. In this proof we supress the dependence on f in the notation. The Gaussian kernel K satisfies $\partial_z K(z) = -zK(z)$. From (4.2) and (4.4) we have $\tilde{L}_n = \bar{y}_2 - \bar{y}_1^2$ where $\bar{y}_l = \sum_i w_i \tilde{Y}_i^l$, $l = 1, 2$. We let $K_j = K(d_j(\mathbf{V}, \mathbf{s}_0)/h_n)$, suppress the dependence on $\mathbf{V}$ and $\mathbf{s}_0$ and write $w_i = K_i / \sum_j K_j$. Then, $\nabla K_i = (-1/h_n^2) K_i d_i \nabla d_i$ and $\nabla w_i = -\left(K_i d_i \nabla d_i (\sum_j K_j) - K_i \sum_j K_j d_j \nabla d_j\right) / (h_n \sum_j K_j)^2$. Next,

$$\begin{aligned}
\nabla \bar{y}_l &= -\frac{1}{h_n^2} \sum_i \tilde{Y}_i^l \frac{K_i d_i \nabla d_i - K_i(\sum_j K_j d_j \nabla d_j)}{(\sum_j K_j)^2} \\
&= -\frac{1}{h_n^2} \sum_i \tilde{Y}_i^l w_i \left(d_i \nabla d_i - \sum_j w_j d_j \nabla d_j \right) \\
&= -\frac{1}{h_n^2} \left(\sum_i \tilde{Y}_i^l w_i d_i \nabla d_i - \sum_j \tilde{Y}_j^l w_j \sum_i w_i d_i \nabla d_i \right) \\
&= -\frac{1}{h_n^2} \sum_i (\tilde{Y}_i^l - \bar{y}_l) w_i d_i \nabla d_i \qquad \text{(B.33)}
\end{aligned}$$

Then, $\nabla \tilde{L}_n = \nabla \bar{y}_2 - 2\bar{y}_1 \nabla \bar{y}_1$, and inserting $\nabla \bar{y}_l$ from (B.33) yields

$$\begin{aligned}
\nabla \tilde{L}_n &= (-1/h_n^2) \sum_i (Y_i^2 - \bar{y}_2 - 2\bar{y}_1 (Y_i - \bar{y}_1)) w_i d_i \nabla d_i \\
&= (1/h_n^2) \left(\sum_i \left(\tilde{L}_n - (Y_i - \bar{y}_1)^2 \right) w_i d_i \nabla d_i \right),
\end{aligned}$$

since $Y_i^2 - \bar{y}_2 - 2\bar{y}_1(Y_i - \bar{y}_1) = (Y_i - \bar{y}_1)^2 - \tilde{L}_n$. □

Acknowledgments

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

References

[1] Adragni, K. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367** 4385–4405. MR2546393

[2] Amemiya, T. (1985). *Advanced Econometrics*. Harvard university press.

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

[4] Boothby, W. M. (2002). *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press. MR0426007

[5] Cleveland, W. S. and Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* **83** 596-610. MR0556476

[6] Cook, R. D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. Wiley, New York. MR1645673

[7] Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statist. Sci.* **22** 1–26. MR2408655

[8] Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** 455–474. MR1902895

[9] Faden, A. M. (1985). The Existence of Regular Conditional Probabilities: Necessary and Sufficient Conditions. *The Annals of Probability* **13** 288–298. MR0770643

[10] Fertl, L. and Bura, E. (2021). Conditional Variance Estimator for Sufficient Dimension Reduction.

[11] Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics* **19** 1–67. MR1091842

[12] Griffiths, P. and Harris, J. (1994). *Principles of algebraic geometry. Wiley Classics Library*. John Wiley & Sons, Inc., New York Reprint of the 1978 original. MR1288523

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

[14] Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

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

[16] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. MR3100153

[17] Jennrich, R. I. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *Ann. Math. Statist.* **40** 633–643. MR0238419

[18] Karr, A. F. (1993). *Probability. Springer Texts in Statistics*. Springer-Verlag New York. MR1231974

[19] Leao Jr., D., Fragoso, M. and Ruffino, P. (2004). Regular Conditional Probability, Disintegration of Probability and Radon Spaces. *Proyecciones (Antofagasta)* **23** 15 – 29. MR2060837

[20] Li, B. (2018). *Sufficient dimension reduction: methods and applications with R*. CRC Press, Taylor & Francis Group. MR3838449

[21] Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **86** 316-327. MR1137117

[22] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review* **81** 134–150. MR3047506

[23] Mickey, M. R., Mundle, P. B., Walker, D. N., Glinski, A. M., C-E-I-R Inc. and Aerospace Research Laboratories (U. S.) (1963). *Test Criteria for Pearson Type III Distributions. ARL (Aerospace Research Laboratories (U.S.))*. Aerospace Research Laboratories, Office of Aerospace Research, United States Air Force.

[24] Tagare, H. D. (2011). Notes on Optimization on Stiefel Manifolds.

[25] Wang, H. and Xia, Y. (2008). Sliced Regression for Dimension Reduction. *Journal of the American Statistical Association* **103** 811-821. MR2524332

[26] Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 363–410. MR1924297

[27] Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39** 3392–3416. MR3012413

[28] Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a Multiple-index regression. *Journal of Multivariate Analysis* **99** 1733-1757. MR2444817

[29] Zeng, P. and Zhu, Y. (2010). An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis* **101** 271 – 290. MR2557633