Ensemble Conditional Variance Estimator for Sufficient Dimension Reduction

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

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

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. It 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. It is shown to outperform *central subspace mean average variance estimation* (csMAVE), its main competitor, under several simulation settings and in a benchmark data set analysis.

1 Introduction

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let Y be 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

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

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.

[ZZ10, Thm. 1] showed that if $(Y, \mathbf{X}^T)^T$ has a joint continuous distribution, (1) is equivalent to

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T \mathbf{X}, \tag{2}$$

where the symbol $\perp\!\!\!\perp$ indicates stochastic independence. The matrix $\mathbf{B}$ is not unique. It can be replaced by any basis of its column space, $\mathrm{span}\{\mathbf{B}\}$. 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 subspaces $\mathcal{S}$ are called dimension reduction subspaces and their intersection, provided it satisfies the conditional independence condition (2), is called the central subspace and denoted by $\mathcal{S}_{Y|\mathbf{X}}$ [see [Coo98, p. 105], [Coo07]].

By their equivalence, under both models (1) and (2), $F_{Y|\mathbf{X}}(y) = F_{Y|\mathbf{B}^T\mathbf{X}}(y)$ and $\mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}\{\mathbf{B}\}$. 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 it can replace $\mathbf{X}$ without any loss of information.

---

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

If the error term in model (1) is additive with $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$, (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 $\mathcal{S} = \mathrm{span}\{\mathbf{B}\}$. The mean subspace, denoted by $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, is the intersection of all subspaces $\mathcal{S}$ such that $\mathbb{E}(Y \mid \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{P}_{\mathcal{S}}\mathbf{X})$ [CL02]. In this case, (1) becomes the classic mean subspace model with $\mathrm{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. [CL02] 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 ([AC09, MZ13, Li18, XTLZ02]). *Linear* refers to the reduction being a linear transformation of the predictor vector. *Minimum Average Variance Estimation* (MAVE) [XTLZ02] is the most competitive and accurate method 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, [Li91]). 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 [CD88].

The *central subspace mean average variance estimation* (csMAVE) [WX08, WY19] is the extension of MAVE that consistently and exhaustively estimates the $\mathrm{span}\{\mathbf{B}\}$ in model (1) without restrictive assumptions limiting its applicability. csMAVE has remained the gold standard since it was proposed by [WX08]. It is based on repeatedly applying MAVE on the sliced target variables $\tilde{f}_u(Y) = \mathbf{1}_{\{s_{u-1} < Y \leq s_u\}}$ for $s_1 < \ldots < s_H$. [WX08] 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 central mean to the central subspace [see [YL11] for a list of references]. [YL11] recognized that these approaches pointed to the same direction: if one can estimate the central 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 [YL11], who also proposed and studied such functional families [see also [Li18] for an overview].

In this paper, we extend the *conditional variance estimator* (CVE) [FB21] to the exhaustive *ensemble conditional variance estimator* for recovering fully the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. *Conditional variance estimation* is a semi-parametric method for the estimation of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ consistently under minimal regularity assumptions on the distribution of $(Y, \mathbf{X}^T)^T$. In contrast to other SDR approaches, it operates by identifying the orthogonal complement of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. In this paper we apply the *conditional variance estimator* (CVE) to identify the mean subspace $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ of transformed responses $f_t(Y)$, where f_t are elements of an *ensemble* $\mathcal{F} = \{f_t : t \in \Omega_T\}$, and then combine them to form the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$.

The paper is organized as follows. In Section 2 we define the notation and concepts we use throughout the paper. A short overview of ensembles is given in Section 2.1. The *ensemble conditional variance estimator* (ECVE) is introduced in Section 3 and the estimation procedure in Section 4. In Section 5, the consistency of the *ensemble conditional variance estimator* for the central subspace is shown. We assess and compare the performance of the estimator vis-a-vis csMAVE via simulations in Section 6 and by applying it to 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 matrix $\mathbf{A}$, $\|\mathbf{A}\|$ denotes its Frobenius norm, and $\|\boldsymbol{a}\|$ the Euclidean norm for a vector $\boldsymbol{a}$. Scalar product refers to the usual Euclidean scalar 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 $q \leq p$,

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

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 [Boo02] and Section 2.1 of [Tag11]]. The set

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

denotes a Grassmann manifold [GH94] 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 Ensembles

[YL11] introduced *ensembles* as a device to extend mean subspace to central subspace SDR methods. The *ensemble* approach of combining mean subspaces to span the central subspace comprizes 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, [YL11] use the term *parametric* ensemble, which we define next.

Definition. *A family $\mathcal{F}$ of measurable functions from $\mathbb{R}$ to $\mathbb{R}$ is called an ensemble. If $\mathcal{F}$ is a family of measurable functions with respect to an index set Ω_T; i.e. $\mathcal{F} = \{f_t : t \in \Omega_T\}$, $\mathcal{F}$ is called a parametric ensemble.*

Let $\mathcal{F}$ be an ensemble, $f \in \mathcal{F}$ and let $f(Y)$, for Y following model (1). 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 [Coo98] or [CL02]].

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

$$\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\} = \mathcal{S}_{Y|\mathbf{X}} \tag{5}$$

As an example, the parametric 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 parametric 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 parametric ensembles $\mathcal{F}$, and associated regularity conditions, that can characterize $\mathcal{S}_{Y|\mathbf{X}}$ from [YL11] next.

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

Indicator ensemble $\mathcal{F} = \{1_{\{z \leq t\}} : t \in \mathbb{R}\}$, 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 $\mathcal{F} = \{h^{-1}K((z-t)/h) : t \in \mathbb{R}, h > 0\}$, 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 determine the distribution uniquely. 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 $\mathcal{S}_{Y|\mathbf{X}}$. [YL11] used the ensemble device to extend `MAVE` [XTLZ02], which targets the mean subspace, to its ensemble version that also estimates the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ consistently.

Theorem 1 [YL11, Thm 2.1] establishes when an ensemble $\mathcal{F}$ is rich enough to characterize $\mathcal{S}_{Y|\mathbf{X}}$.

Theorem 1. *Let $\mathcal{B} = \{1_A : A \text{ is a Borel set in } supp(Y)\}$ 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 distribution F_Y of the response 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 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. *If a parametric ensemble $\mathcal{F}$ characterizes $\mathcal{S}_{Y|\mathbf{X}}$, then there exist finitely many functions $f_t \in \mathcal{F}$, with $t = 1, \ldots, m$ and $m \in \mathbb{N}$, such that*

$$\text{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : t \in 1, \ldots, m\} = \mathcal{S}_{Y|\mathbf{X}}$$

3

Proof: Let $k = \dim(\mathcal{S}_{Y|\mathbf{X}}) \leq p$. Since $\mathcal{F}$ characterizes $\mathcal{S}_{Y|\mathbf{X}}$, $\dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}) = k_t \leq k$ by (5) for any t. If $k_t = 0$, $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \{\mathbf{0}\}$ so the corresponding f_t does not contribute to (5). Assume $k_t \geq 1$. If there were infinitely many $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \neq \{\mathbf{0}\}$ of dimension at least 1, whose span is $\mathcal{S}_{Y|\mathbf{X}}$, then infinitely as many are identical, otherwise the dimension of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ would be infinite, contradicting that $\dim(\mathcal{S}_{Y|\mathbf{X}}) = k < \infty$. $\qquad \square$

The importance of Theorem 2 lies in the fact that the search to characterize the central subspace is over a finite set, even though it does not offer tools for identifying the elements of the ensemble.

3   Ensemble CVE

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

(E.1). *Model* (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,* $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathcal{S}(p, k)$, $\mathbf{X} \in \mathbb{R}^p$ *is 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* $\mathbb{V}ar(\mathbf{X}) = \mathbf{\Sigma}_\mathbf{x}$ *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). *For a parametric ensemble* $\mathcal{F}$, *its index set* Ω_T *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 an ensemble* $\mathcal{F}$ *we assume that for all* $f \in \mathcal{F}$, *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 $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$. 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 assumption for $supp(f_\mathbf{X})$ is not as restrictive as it might seem. [YLC08, 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) simply states 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 a parametric 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 are twice continuously differentiable. In this case, the $8th$ moments exist since the complex exponential and indicator functions are bounded.

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

$$\tilde{L}_\mathcal{F}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}ar\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) \tag{6}$$

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

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

$$L_\mathcal{F}(\mathbf{V}) = \int_{\Omega_T} \int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}, f_t) dF_\mathbf{X}(\mathbf{x}) dF_T(t) \tag{7}$$

$$= \mathbb{E}_{t \sim F_T}\left(\mathbb{E}_\mathbf{X}\left(\tilde{L}_\mathcal{F}(\mathbf{V}, \mathbf{X}, f_t)\right)\right) = \mathbb{E}_{t \sim F_T}(L^*_\mathcal{F}(\mathbf{V}, f_t)),$$

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

$$L^*_\mathcal{F}(\mathbf{V}, f_t) = \mathbb{E}_\mathbf{X}\left(\tilde{L}_\mathcal{F}(\mathbf{V}, \mathbf{X}, f_t)\right). \tag{8}$$

For the identity function, $f_{t_0}(z) = z$, (8) is the target function of the *conditional variance estimation* proposed in [FB21]. 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).

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 $\mathbf{X}$.

4

Theorem 3. *Assume (E.1) and (E.2) hold, in particular model (1) holds. Let $\widetilde{\mathbf{B}}$ be a basis of $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$; i.e.* $\mathrm{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}\{\mathbf{B}\}$. *Then, for any $f \in \mathcal{F}$ for which assumption (E.4) holds,*

$$f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}, \tag{9}$$

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 3, any response Y can be written as an additive error via the decomposition (9). The predictors and the additive error term are only required to be conditionally uncorrelated in model (9). The *conditional variance estimator* [FB21] also estimated $\widetilde{\mathbf{B}}$ in (9) but under the more restrictive condition of predictor and error independence.

Proof of Theorem 3.

$$f(Y) = \mathbb{E}\left(f(Y) \mid \mathbf{X}\right) + \underbrace{f(Y) - \mathbb{E}\left(f(Y) \mid \mathbf{X}\right)}_{\tilde{\epsilon}} = \mathbb{E}\left(f(Y) \mid \mathbf{X}\right) + \tilde{\epsilon}$$

$$= \mathbb{E}\left(f(Y)|\widetilde{\mathbf{B}}^T\mathbf{X}\right) + \tilde{\epsilon} = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}$$

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 4. *Assume (E.1) and (E.2) hold. Let $\mathcal{F}$ be a parametric ensemble, $\mathbf{s}_0 \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$, $\mathbf{V} \in S(p,q)$ defined in (3). Then, for any $f \in \mathcal{F}$ for which assumption (E.4) holds,*

$$\tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{s}_0, f) = \mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1^2(\mathbf{V}, \mathbf{s}_0, f) + \mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) \tag{10}$$

where

$$\mu_l(\mathbf{V}, \mathbf{s}_0, f) = \int_{\mathbb{R}^q} g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1)^l \frac{f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} d\mathbf{r}_1 = \frac{t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)}, \tag{11}$$

for g given in (9) with

$$t^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = \int_{\mathbb{R}^q} g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1, \tag{12}$$

and

$$\mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\})$$

$$= \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1 / \int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r} = \frac{\tilde{h}(\mathbf{V}, \mathbf{s}_0, f)}{t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)} \tag{13}$$

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 $L_{\mathcal{F}}^*(\mathbf{V}, f_t)$ in (8) is well defined and continuous,

$$\mathbf{V}_q^t = \mathrm{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L_{\mathcal{F}}^*(\mathbf{V}, f_t) \tag{14}$$

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

$$\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \mathrm{span}\{\mathbf{V}_q^t\}^{\perp}. \tag{15}$$

[FB21] assumed model $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$ with $\epsilon \perp\!\!\!\perp \mathbf{X}$, which implies $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathrm{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$. [FB21] showed that the *conditional variance estimator (CVE)* can identify $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ at the population level.

Theorem 4 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 (14) is not unique since for all orthogonal $\mathbf{O} \in \mathbb{R}^{q \times q}$, $L_{\mathcal{F}}^*(\mathbf{V}_q^t\mathbf{O}, f_t) = L_{\mathcal{F}}^*(\mathbf{V}_q^t, f_t)$ as $L_{\mathcal{F}}^*(\mathbf{V}_q^t, f_t)$ depends on $\mathbf{V}_q^t$ only through $\mathrm{span}\{\mathbf{V}_q^t\}$ by (6). Nevertheless, it is a unique minimizer over the Grassmann manifold $Gr(p,q)$ in

(4). To see this, suppose $\mathbf{V} \in \mathcal{S}(p, q)$ is an arbitrary basis of a subspace $\mathbf{M} \in Gr(p, q)$. We can identify $\mathbf{M}$ through the projection $\mathbf{P_M} = \mathbf{VV}^T$. By (31), we write $\mathbf{x} = \mathbf{Vr}_1 + \mathbf{Ur}_2$. Application of the Fubini-Tonelli Theorem yields

$$\tilde{t}^{(l)}(\mathbf{P_M}, \mathbf{s}_0, f) = \int_{\text{supp}(f_{\mathbf{x}})} g(\mathbf{B}^T \mathbf{s}_0 + \mathbf{B}^T \mathbf{P_M} \mathbf{x})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{P_M} \mathbf{x}) d\mathbf{x} \tag{16}$$

$$= t^{(l)}(\mathbf{V}, \mathbf{s}_0, f) \int_{\text{supp}(f_{\mathbf{x}}) \cap \mathbb{R}^{p-q}} d\mathbf{r}_2.$$

Therefore $\tilde{t}^{(l)}(\mathbf{P_M}, \mathbf{s}_0, f)/\tilde{t}^{(0)}(\mathbf{P_M}, \mathbf{s}_0, f) = t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)/t^{(0)}(\mathbf{V}, \mathbf{s}_0, f)$ and $\mu_l(\cdot, \mathbf{s}_0, f)$ in (11) can also be viewed as a function from $Gr(p, q)$ to $\mathbb{R}$.

Next we define the *ensemble conditional variance estimator (ECVE)* for a parametric ensemble $\mathcal{F}$ which characterizes the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. Following the *ensemble minimum average variance estimation* formulation in [YL11], we extend the original objective function by integrating over the index random variable $t \sim F_T$ in (7) that indexes the ensemble $\mathcal{F}$ as [YL11].

Definition 5. *Let*

$$\mathbf{V}_q = \text{argmin}_{\mathbf{V} \in S(p,q)} L_{\mathcal{F}}(\mathbf{V}) \tag{17}$$

*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* $\text{span}\{\mathbf{V}_q\}^{\perp}$.

Theorem 6. *Assume (E.1), (E.2), (E.3), and (E.4) hold, and that the function $h(\cdot)$ defined in Theorem 4 is continuous. Let $\mathcal{F}$ be a parametric 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 $S(p, q)$, which is defined in (3), with $q = p - k$. Then, $\mathbf{V}_q$ in (17) is well defined and*

$$\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{V}_q\}^{\perp}. \tag{18}$$

4 Estimation of the ensemble CVE

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

$$d_i(\mathbf{V}, \mathbf{s}_0) = \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0 + \text{span}\{\mathbf{V}\}} \mathbf{X}_i\|_2^2 = \|\mathbf{X}_i - \mathbf{s}_0\|_2^2 - \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{VV}^{\top}(\mathbf{X}_i - \mathbf{s}_0)\rangle$$

$$= \|(\mathbf{I}_p - \mathbf{VV}^{\top})(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 = \|\mathbf{Q_V}(\mathbf{X}_i - \mathbf{s}_0)\|_2^2 \tag{19}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P_V} = \mathbf{VV}^{\top}$ and $\mathbf{Q_V} = \mathbf{I}_p - \mathbf{P_V}$. 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 subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$. In order to obtain pointwise convergence for the ensemble CVE, we use the following bias and variance assumptions on the bandwidth, as typical in nonparametric estimation.

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

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

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

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

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

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

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

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

The Gaussian kernel $K(z) = \exp(-z^2)$, for example, fulfills both (K.1) and (K.2) [see [Han08]], and will be used throughout the paper.

For $i = 1, \ldots, n$, we let

$$w_i(\mathbf{V}, \mathbf{s}_0) = \frac{K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{d_j(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)} \tag{20}$$

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0, f) = \sum_{i=1}^{n} w_i(\mathbf{V}, \mathbf{s}_0) f(Y_i)^l \quad \text{for} \quad l = 1, 2 \tag{21}$$

We estimate $\tilde{L}_{\mathcal{F}}(\mathbf{V}, s_0, f)$ in (10) with

$$\tilde{L}_{n,\mathcal{F}}(\mathbf{V}, s_0, f) = \bar{y}_2(\mathbf{V}, \mathbf{s}_0, f) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0, f)^2, \tag{22}$$

and the objective function $L_{\mathcal{F}}^*(\mathbf{V}, f)$ in (8) with

$$L_n^*(\mathbf{V}, f) = \frac{1}{n} \sum_{i=1}^{n} \tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f), \tag{23}$$

where each data point $\mathbf{X}_i$ is a shifting point. For a parametric 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 $\lim_{n\to\infty} m_n = \infty$, the final estimate of the objective function in (7) is given by

$$L_{n,\mathcal{F}}(\mathbf{V}) = \frac{1}{m_n} \sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) \tag{24}$$

The ensemble conditional variance estimator (ECVE) is defined to be any basis of $\text{span}\{\hat{\mathbf{V}}_q\}^{\perp}$, where

$$\hat{\mathbf{V}}_q = \text{argmin}_{\mathbf{V} \in S(p,q)} L_{n,\mathcal{F}}(\mathbf{V}) \tag{25}$$

We use the same algorithm as in [FB21] to solve the optimization problem (25). It requires the explicit form of the gradient of (24). Theorem 7 provides the gradient when a Gaussian kernel is used.

Theorem 7. *The gradient of $\tilde{L}_{n,\mathcal{F}}(\mathbf{V}, s_0, f)$ in (22) is given by*

$$\nabla_{\mathbf{V}} \tilde{L}_{n,\mathcal{F}}(\mathbf{V}, s_0, f) = \frac{1}{h_n^2} \sum_{i=1}^{n} (\tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{s}_0, f) - (f(Y_i) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0, f))^2) w_i d_i \nabla_{\mathbf{V}} d_i(\mathbf{V}, \mathbf{s}_0) \in \mathbb{R}^{p \times q},$$

and the gradient of $L_{n,\mathcal{F}}(\mathbf{V})$ in (24) is

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

In the implementation of ECVE, we follow [FB21] and set the bandwidth to

$$h_n = 1.2^2 \frac{2\text{tr}(\widehat{\Sigma}_{\mathbf{x}})}{p} \left(n^{-1/(4+p-q)} \right)^2. \tag{26}$$

where $\widehat{\Sigma}_{\mathbf{x}} = (1/n) \sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$ and $\bar{\mathbf{X}} = (1/n) \sum_i \mathbf{X}_i$.

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

The set of points $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$ represents a *slice* in the subspace of $\mathbb{R}^p$ about $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$. In the estimation of $L(\mathbf{V})$ two different weighting schemes are used: (a) *Within slices*: The weights are defined in (20) and are used to calculate (22). (b) *Between slices*: Equal weights $1/n$ are used to calculate (23). Another idea for the between slices weighting is to assign more weight to slices with more points. This can be realized by altering (23) to

$$L_n^{(w)}(\mathbf{V}, f) = \sum_{i=1}^{n} \tilde{w}(\mathbf{V}, \mathbf{X}_i) \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f), \quad \text{with} \tag{27}$$

$$\tilde{w}(\mathbf{V}, \mathbf{X}_i) = \frac{\sum_{j=1}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n) - 1}{\sum_{l,u=1}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n) - n} = \frac{\sum_{j=1,j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)}{\sum_{l,u=1,l\neq u}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n)} \tag{28}$$

The denominator in (28) guarantees the weights $\tilde{w}(\mathbf{V}, \mathbf{X}_i)$ sum up to one. If (27) instead of (23) is used in (24) we refer to this method as *weighted ensemble conditional variance estimation*.

For example, if a rectangular kernel is used, $\sum_{j=1,j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$ is 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 the ECVE

The consistency of ECVE derives from the consistency of CVE [FB21] that targets a specific $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ and the fact that we can recover $\mathcal{S}_{Y|\mathbf{X}}$ from $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ across all transformations $f_t \in \mathcal{F} = \{f_t : t \in \Omega_T\}$ for an ensemble that characterizes $\mathcal{S}_{Y|\mathbf{X}}$. This is achieved in sequential steps from Theorem 8, which is the main building block, to Theorem 11. The proofs are technical and lengthy, and, thus, are given in the Appendix.

Theorem 8. *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 a parametric 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 3. Then, $L_n^*(\mathbf{V}, f)$, defined in (23), converges uniformly in probability to $L^*(\mathbf{V}, f)$ in (8) for all $f \in \mathcal{F}$; i.e.,*

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

Next, Theorem 9 shows that ensemble conditional variance estimator is consistent for $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ for any transformation f.

Theorem 9. *Under the same conditions as Theorem 8, the conditional variance estimator $\text{span}\{\widehat{\mathbf{B}}_{k_t}^t\}$ estimates $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ consistently, for $f_t \in \mathcal{F}$. That is,*

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

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

$$\widehat{\mathbf{V}}_{k_t}^t = \text{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L_{n,\mathcal{F}}^*(\mathbf{V}, f_t).$$

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

A straightforward application of Theorem 9, using the identity function, obtains that $\mathcal{S}_{E(Y|\mathbf{X})}$ can be consistently estimated by ECVE.

Theorem 10. *Assume the conditions of Theorem 8 hold. Let $\mathcal{F}$ be a parametric 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 $L_{n,\mathcal{F}}(\mathbf{V})$, defined in (24), converges uniformly in probability to $L_{\mathcal{F}}(\mathbf{V})$ in (7); i.e.,*

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

The assumption $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$ in Theorem 10 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 8 can be dropped since obviously no truncation is needed.

The rate of convergence of m_n is not characterized in Theorem 10. In the simulation studies of Sections 6.2 and 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 the ensemble CVE is shown in Theorem 11.

Theorem 11. *Assume the conditions of Theorem 8 and (E.3) hold. Let $\mathcal{F}$ be a parametric 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 $\text{span}\{\widehat{\mathbf{V}}_q\}^\perp$, where $\widehat{\mathbf{V}}_q$ is defined in (25) with $q = p - k$ and $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$,*

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

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 this section we study the influence of the number of functions of the ensemble $\mathcal{F}$, m_n in (24), on the accuracy of the ensemble conditional variance estimation. In Theorem 10 and 11, how fast m_n approaches ∞ is unspecified. We consider the 2-dimensional regression model

$$Y = (\mathbf{b}_2^T \mathbf{X}) + (0.5 + (\mathbf{b}_1^T \mathbf{X})^2)\epsilon, \tag{29}$$

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, $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{b}_2\} \subsetneq \mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$, with $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)$.

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.

For each ensemble, we form the ensemble conditional variance estimator and its weighted version as in section 4.1, see also [FB21]. The results of 100 replications for each method and each m are displayed in Figure 1. We assess the estimation accuracy with $\text{err}_{j,m} = \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|/(2k)^{1/2}$, $j = 1, \ldots, 100$, $m \in \{2, 4, 8, 10, 26, 50, 76, 100\}$. ECVE's main competitor, csMAVE, which does not vary with m, estimate of the central subspace has median error 0.2 with a wide range from 0.1 to 0.6. The estimation accuracy of Fourier, Indicator and Box-Cox ECVE vary over m and is on par or better for some m values.

For the Fourier basis, fewer basis functions give the best performance, the indicator and BoxCox 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 a choice of a small m number of basis functions. Based on this and further unreported simulations, we set the default value of m to

$$m_n = \begin{cases} \lceil \log(n) \rceil, \text{if} & \lceil \log(n) \rceil \quad \text{even} \\ \lceil \log(n) \rceil + 1, \text{if} & \lceil \log(n) \rceil \quad \text{odd} \end{cases} \tag{30}$$

for all simulations in Section 6.2, 6.3 and the data analysis in Section 7.

6.2 Demonstrating consistency

We explore the consistency rate of the *conditional variance estimator (CVE)* and *ensemble conditional variance estimator (CVE)*, csMAVE and mMAVE in model (29).

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 [FB21] and mMAVE (VII) in [XTLZ02].

The simulation is performed as follows. We generate 100 i.i.d samples $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$ from (29) for each sample size $n = 100, 200, 400, 600, 800, 1000$. Model (29) 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 2 displays the distribution of $err_{j,n}$ for increasing 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 11, in contrast to the Indicator and Fourier ensembles. The Fourier, Indicator ECVE and csMAVE estimate $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$ consistently and 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 accuracy

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 $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{Z} \in \mathbb{R}^p$, where $\boldsymbol{\Sigma} = (\Sigma_{ij})_{i,j=1,\ldots,p}$, $\Sigma_{i,j} = 0.5^{|i-j|}$. 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 [WX08], 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* (ECVE). The monomial and BoxCox 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
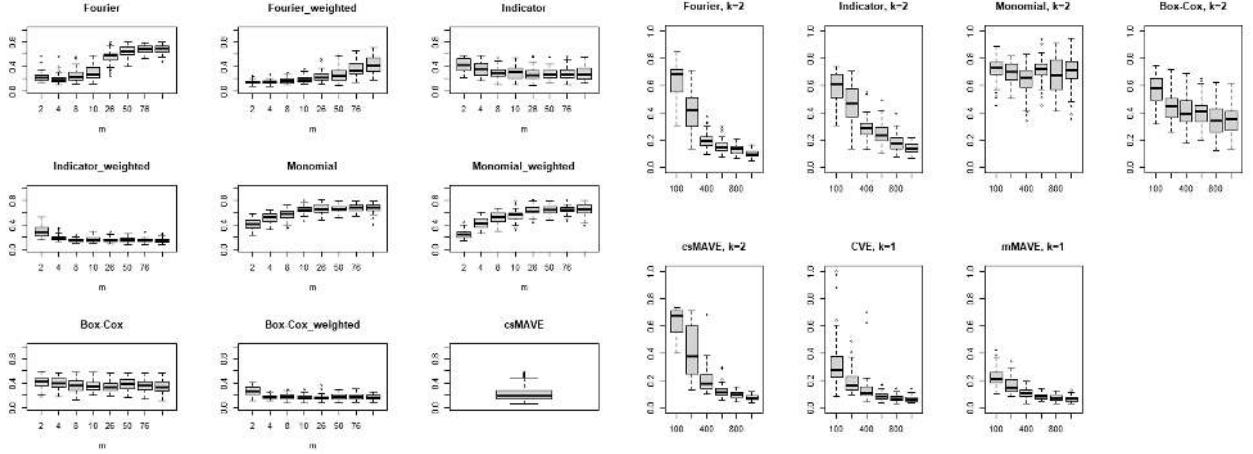
Figure 1: Box plots of the estimation errors over 100 replications of model (29) with $n = 300$ over $m = |\mathcal{F}| = (2, 4, 8, 10, 26, 50, 76, 100)$ across four ensembles.

Figure 2: Estimation error distribution of model (29) plotted over $n = (100, 200, 400, 600, 800, 1000)$ for the seven (I-VII) methods

against the reference method `csMAVE` [WX08], which is implemented in the R package `MAVE`. The source code for *conditional variance estimation* and its ensemble version is available at `https://git.art-ist.cc/daniel/CVE`.

Table 1: Models

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	$Y = \frac{\mathbf{b}_1^T \mathbf{X}}{0.5 + (1.5 + \mathbf{b}_2^T \mathbf{X})^2} + (|\mathbf{b}_1^T \mathbf{X}| + (\mathbf{b}_2^T \mathbf{X})^2 + 0.5)\epsilon$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	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 with the specified distribution of $\mathbf{X}$ and sample size n. We estimate $\mathbf{B}$ using the four ECVE methods and csMAVE. The accuracy of the estimates is assessed using $err = \|\mathbf{P}_\mathbf{B} - \mathbf{P}_{\widehat{\mathbf{B}}}\|_2 / \sqrt{2k} \in [0, 1]$, where $\mathbf{P}_\mathbf{B} = \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1}\mathbf{B}^T$ is the orthogonal projection matrix on $\text{span}\{\mathbf{B}\}$. The factor $\sqrt{2k}$ normalizes the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement. The results are displayed in Tables 2-8. In M1, which is taken from [WX08], the mean subspace agrees with the central subspace, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$, but due to the unboundedness of the link function $g(x) = 1/x$ most mean subspace estimation methods, such as SIR, mean MAVE and CVE, fail. In contrast, all 4 ensemble CVE methods and csMAVE succeed in identifying the minimal dimension reduction subspace, with ensemble CVE performing slightly better, as can be seen in Table 2. In particular, `Fourier` is the best performing method. M2, is a two dimensional mean subspace model, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$, and in Table 3 we see that `csMAVE` is the best performing method. M3 is the same as model (29) and here 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 determines also the heteroskedasticity, agrees with the central subspace. In Table 5 we see 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, the heteroskedasticity is induced by an interaction term, and the three dimensional central subspace model is a proper superset of the one dimensional mean subspace. In Table 6 we see that M5 is quite challenging for all five methods, therefore we increase the sample size n to 800. For M5, the two weighted ensemble conditional variance estimators are the best performing methods followed by `csMAVE`.

M6 is a one dimensional pure central subspace model, whereas the mean subspace is 0. In Table 7, we see that for $n = 100$ the two weighted ECVEs are the best performing methods and for higher sample sizes csMAVE is slightly more accurate than the ECVE methods.

In M7 the one dimensional mean subspace agrees with the central subspace, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$, and the conditional first and second moments, $\mathbb{E}(Y^l \mid \mathbf{X})$ for $l = 1, 2$, are highly nonlinear and periodic functions of the sufficient reduction. In Table 8, we see that all ensemble conditional variance estimators clearly outperform csMAVE.

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

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
I	100	**0.172** (0.047)	0.201 (0.054)	0.248 (0.064)	0.265 (0.063)	0.210 (0.063)
I	200	**0.120** (0.029)	0.142 (0.037)	0.182 (0.045)	0.197 (0.049)	0.128 (0.037)
I	400	**0.079** (0.020)	0.091 (0.024)	0.126 (0.037)	0.136 (0.040)	0.080 (0.024)
II	100	**0.174** (0.038)	0.196 (0.049)	0.241 (0.055)	0.254 (0.056)	0.193 (0.059)
II	200	**0.110** (0.031)	0.127 (0.033)	0.170 (0.043)	0.182 (0.045)	0.121 (0.036)
II	400	**0.078** (0.021)	0.091 (0.026)	0.122 (0.031)	0.132 (0.033)	0.079 (0.020)
III	100	**0.187** (0.045)	0.218 (0.053)	0.256 (0.060)	0.263 (0.058)	0.204 (0.066)
III	200	**0.118** (0.031)	0.137 (0.038)	0.171 (0.043)	0.179 (0.042)	**0.118** (0.033)
III	400	0.082 (0.020)	0.101 (0.029)	0.127 (0.031)	0.132 (0.032)	**0.079** (0.022)

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

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
I	100	0.670 (0.089)	0.601 (0.135)	0.629 (0.130)	0.582 (0.140)	**0.575** (0.176)
I	200	0.478 (0.201)	0.388 (0.152)	0.436 (0.193)	0.407 (0.162)	**0.219** (0.136)
I	400	0.226 (0.153)	0.201 (0.074)	0.231 (0.127)	0.236 (0.111)	**0.098** (0.025)
II	100	0.663 (0.097)	0.652 (0.104)	0.687 (0.057)	0.658 (0.080)	**0.544** (0.176)
II	200	0.525 (0.171)	0.468 (0.171)	0.601 (0.127)	0.539 (0.148)	**0.182** (0.096)
II	400	0.267 (0.081)	0.307 (0.146)	0.375 (0.154)	0.357 (0.141)	**0.087** (0.021)
III	100	0.657 (0.104)	0.590 (0.148)	**0.530** (0.155)	0.542 (0.148)	0.603 (0.193)
III	200	0.421 (0.203)	0.367 (0.165)	0.306 (0.147)	0.336 (0.151)	**0.240** (0.193)
III	400	0.170 (0.110)	0.170 (0.071)	0.144 (0.053)	0.170 (0.063)	**0.089** (0.019)

7 Boston Housing Data

We apply the ensemble conditional variance estimator and csMAVE to the Boston Housing data set. This data set has been extensively used as a benchmark for assessing regression methods [see, for example, [JWHT13]], 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 prior to 1940), dis (weighted distances to five Boston employment

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

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
I	100	0.744 (0.056)	0.657 (0.113)	0.668 (0.083)	**0.561** (0.142)	0.602 (0.147)
I	200	0.702 (0.061)	0.472 (0.177)	0.559 (0.147)	**0.369** (0.155)	0.374 (0.148)
I	400	0.621 (0.148)	0.252 (0.102)	0.408 (0.177)	0.223 (0.064)	**0.203** (0.061)
II	100	0.751 (0.041)	0.698 (0.076)	0.683 (0.080)	**0.570** (0.136)	0.635 (0.136)
II	200	0.719 (0.040)	0.521 (0.163)	0.584 (0.111)	**0.355** (0.097)	0.387 (0.144)
II	400	0.686 (0.079)	0.267 (0.084)	0.452 (0.153)	0.252 (0.052)	**0.201** (0.045)
III	100	0.739 (0.073)	0.676 (0.106)	0.654 (0.105)	**0.563** (0.150)	0.571 (0.120)
III	200	0.704 (0.048)	0.546 (0.162)	0.523 (0.171)	0.368 (0.153)	**0.330** (0.131)
III	400	0.616 (0.151)	0.252 (0.113)	0.297 (0.106)	0.202 (0.055)	**0.179** (0.042)

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

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

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 [Fri91], 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 results. 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 `ensemble CVE` and `csMAVE` are approximately on par with the variable selection methods in [JWHT13, Section 8.3.3].

Moreover, we plot the standardized response `medv` against the reduced `Fourier` and `csMAVE` predictors, $\mathbf{B}^T \mathbf{X}$, in Figure 3. 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, $|\widehat{\mathbf{B}}_{\text{Fourier}}^T \widehat{\mathbf{B}}_{\text{csMAVE}}| = 0.786$. The coefficients of the reductions, $\widehat{\mathbf{B}}_{\text{Fourier}}$ and $\widehat{\mathbf{B}}_{\text{csMAVE}}$, are reported in Table 10. For the `Fourier` ensemble, the variables `rm` and `lstat` have the highest influence on the target variable

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

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

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

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
I	100	0.304 (0.092)	**0.294** (0.082)	0.492 (0.135)	0.299 (0.087)	0.539 (0.255)
I	200	0.217 (0.057)	0.213 (0.054)	0.329 (0.107)	0.205 (0.059)	**0.194** (0.061)
I	400	0.142 (0.036)	0.146 (0.035)	0.199 (0.069)	0.138 (0.039)	**0.114** (0.034)
II	100	0.308 (0.094)	**0.293** (0.073)	0.479 (0.129)	0.299 (0.086)	0.488 (0.248)
II	200	0.205 (0.058)	0.210 (0.057)	0.321 (0.095)	0.210 (0.058)	**0.192** (0.061)
II	400	0.144 (0.039)	0.150 (0.042)	0.190 (0.055)	0.142 (0.045)	**0.111** (0.032)
III	100	0.373 (0.152)	0.375 (0.175)	0.504 (0.143)	**0.322** (0.083)	0.562 (0.273)
III	200	0.226 (0.065)	0.230 (0.070)	0.340 (0.100)	**0.218** (0.060)	**0.218** (0.083)
III	400	0.149 (0.039)	0.151 (0.038)	0.194 (0.068)	0.146 (0.042)	**0.114** (0.032)

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

Distribution	n	Fourier	Fourier_weighted	Indicator	Indicator_weighted	csMAVE
I	100	0.273 (0.169)	**0.237** (0.050)	0.241 (0.136)	0.252 (0.158)	0.790 (0.316)
I	200	0.160 (0.093)	0.159 (0.041)	**0.143** (0.083)	0.153 (0.093)	0.425 (0.391)
I	400	0.098 (0.024)	0.104 (0.025)	**0.088** (0.021)	0.102 (0.093)	0.127 (0.202)
II	100	**0.233** (0.057)	0.260 (0.134)	0.236 (0.142)	0.265 (0.185)	0.902 (0.219)
II	200	0.154 (0.058)	0.176 (0.124)	**0.145** (0.093)	0.150 (0.094)	0.649 (0.414)
II	400	0.097 (0.025)	0.110 (0.094)	**0.087** (0.022)	0.099 (0.093)	0.295 (0.391)
III	100	0.274 (0.201)	0.303 (0.237)	**0.238** (0.160)	0.298 (0.242)	0.933 (0.163)
III	200	0.167 (0.120)	0.188 (0.159)	**0.159** (0.150)	0.167 (0.144)	0.678 (0.408)
III	400	0.100 (0.023)	0.116 (0.090)	**0.089** (0.023)	0.112 (0.129)	0.375 (0.431)

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

medv. This agrees with the analysis in [JWHT13, 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 10: Rounded coefficients of the estimated reductions for $\widehat{\mathbf{B}}_{\text{Fourier}}$ and $\widehat{\mathbf{B}}_{\text{csMAVE}}$ 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 [FB21] to the ensemble conditional variance estimation (ECVE), which exhaustively estimates the *central subspace*, by applying the ensemble device introduced by [YL11]. In Section 5 we showed 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 [WX08] is consistent.

We analysed the estimation accuracy of ECVE in Section 6. We found that it is either on par with csMAVE or that it exhibits substantial performance improvement in certain models. We could not characterize the defining features of the models for which the ensemble conditional variance estimation outperforms csMAVE. This is an interesting line of further research together with establishing more theoretical results such as the rate of convergence, estimation of the structural dimension, and the limiting distribution of the estimator.

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.

Throughout, the dimension of the central subspace, $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$, is assumed to be known. The derivation of asymptotic tests for dimension is technically very challenging due to the lack of closed-form solution and the lack of
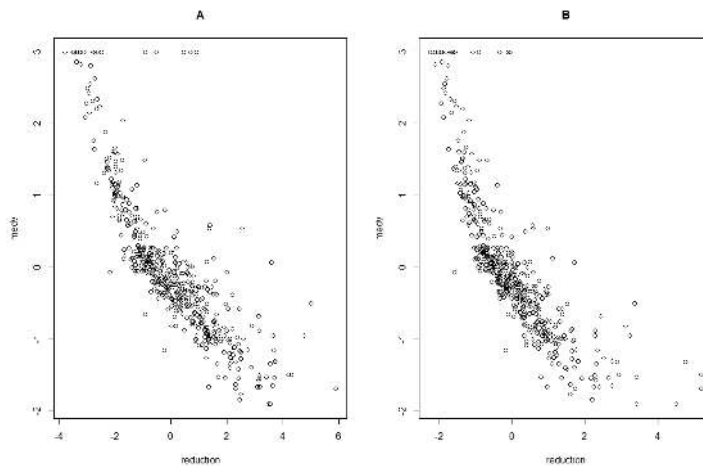
Figure 3: Panel A: Y vs. $\widehat{\mathbf{B}}_{\texttt{Fourier}}^T\mathbf{X}$. Panel B: Y vs. $\widehat{\mathbf{B}}_{\texttt{csMAVE}}^T\mathbf{X}$

independence of all quantities in the calculation. The dimension can be estimated via cross-validation, as in [WX08] and [FB21], or information criteria.

Acknowledgements

References

[AC09]   Kofi P. Adragni and R. Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 11 2009.

[Ame85]   Takeshi Amemiya. *Advanced econometrics*. Harvard university press, 1985.

[Boo02]   W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2002.

[CD88]   William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[CL02]   R.Dennis Cook and Bing Li. Dimension reduction for conditional mean in regression. *Ann. Statist.*, 30(2):455–474, 04 2002.

[Coo98]   Dennis R. Cook. *Regression Graphics: Ideas for studying regressions through graphics*. Wiley, New York, 1998.

[Coo07]   R. Dennis Cook. Fisher lecture: Dimension reduction in regression. *Statist. Sci.*, 22(1):1–26, 02 2007.

[Fad85]   Arnold M. Faden. The existence of regular conditional probabilities: Necessary and sufficient conditions. *The Annals of Probability*, 13(1):288–298, 1985.

[FB21]   Lukas Fertl and Efstathia Bura. Conditional variance estimator for sufficient dimension reduction, 2021.

[Fri91]   Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[GH94]   Phillip Griffiths and Joseph Harris. *Principles of algebraic geometry*. Wiley Classics Library. John Wiley & Sons, Inc., New York, 1994. Reprint of the 1978 original.

[Han08]   Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24:726–748, 2008.

[Heu95]   H. Heuser. *Analysis 2, 9 Auflage*. Teubner, 1995.

[Jen69]   Robert I. Jennrich. Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643, 04 1969.

[JWHT13]   Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[Kar93]   Alan F. Karr. *Probability*. Springer Texts in Statistics. Springer-Verlag New York, 1993.

[Li91]   K. C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

[Li18]   Bing Li. *Sufficient dimension reduction: methods and applications with R*. CRC Press, Taylor & Francis Group, 2018.

[LJFR04]   D. Leao Jr., M. Fragoso, and P. Ruffino. Regular conditional probability, disintegration of probability and radon spaces. *Proyecciones (Antofagasta)*, 23:15 – 29, 05 2004.

[MMW⁺63] M.R. Mickey, P.B. Mundle, D.N. Walker, A.M. Glinski, Inc C-E-I-R, and Aerospace Research Laboratories (U.S.). *Test Criteria for Pearson Type III Distributions*. ARL (Aerospace Research Laboratories (U.S.))). Aerospace Research Laboratories, Office of Aerospace Research, United States Air Force, 1963.

[MZ13]   Yanyuan Ma and Liping Zhu. A review on dimension reduction. *International Statistical Review*, 81(1):134–150, 4 2013.

[S.N27]   S.N.Bernstein. *Theory of Probability*. 1927.

[Tag11]   Hemant D. Tagare. Notes on optimization on stiefel manifolds, January 2011.

[WX08]   Hansheng Wang and Yingcun Xia. Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821, 2008.

[WY19]   Hang Weiqiang and Xia Yingcun. *MAVE: Methods for Dimension Reduction*, 2019. R package version 1.3.10.

[XTLZ02]   Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):363–410, 2002.

[YL11]   Xiangrong Yin and Bing Li. Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.*, 39(6):3392–3416, 12 2011.

[YLC08]   Xiangrong Yin, Bing Li, and R. Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99:1733–1757, 09 2008.

[ZZ10]   Peng Zeng and Yu Zhu. An integral transform method for estimating the central mean and central subspaces. *Journal of Multivariate Analysis*, 101(1):271 – 290, 2010.

Appendix

For any $\mathbf{V} \in \mathcal{S}(p, q)$, defined in (3), we generically denote a basis of the orthogonal complement of its column space span$\{\mathbf{V}\}$, by $\mathbf{U}$. That is, $\mathbf{U} \in \mathcal{S}(p, p-q)$ such that span$\{\mathbf{V}\} \perp$ span$\{\mathbf{U}\}$ and span$\{\mathbf{V}\} \cup$ 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

$$\mathbf{x} = \mathbf{s}_0 + \mathbf{P}_\mathbf{V}(\mathbf{x} - \mathbf{s}_0) + \mathbf{P}_\mathbf{U}(\mathbf{x} - \mathbf{s}_0) = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2 \tag{31}$$

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

Proof of Theorem 4. The density of $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 +$ span$\{\mathbf{V}\}$ is given by

$$f_{\mathbf{X}|\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}}(\mathbf{r}_1) = \frac{f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}} \tag{32}$$

where $\mathbf{X}$ is the p-dimensional continuous random covariate vector with density $f_\mathbf{X}(\mathbf{x})$, $\mathbf{s}_0 \in$ supp$(f_\mathbf{X}) \subset \mathbb{R}^p$, and $\mathbf{V}$ belongs to the Stiefel manifold $\mathcal{S}(p, q)$ defined in (3). Equation (32) follows from Theorem 3.1 of [LJFR04] 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 $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 +$ span$\{\mathbf{V}\}$ [see also [Fad85]]. Further, the measure is concentrated on the affine subspace $\mathbf{s}_0 +$ span$\{\mathbf{V}\} \subset \mathbb{R}^p$ with density (32), which follows from Definition 8.38, Theorem 8.39 of [Kar93], the orthogonal decomposition (31), 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, span$\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} =$ span$\{\mathbf{B}\}$. By Theorem 3, $f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}$, with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and g twice continuously differentiable. Therefore,

$$\tilde{L}_\mathcal{F}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\text{ar}\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right)$$

$$= \mathbb{V}\text{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) + 2\text{cov}\left(\tilde{\epsilon}, g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right)$$

$$+ \mathbb{V}\text{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right)$$

$$= \mathbb{V}\text{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) + \mathbb{V}\text{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) \tag{33}$$

The covariance term in (33) vanishes since

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

$$-E\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) \mathbb{E}\left(\underbrace{\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X})}_{=0} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) = 0$$

i.e. the sigma field generated by $\mathbf{X} \in \mathbf{s}_0 +$ span$\{\mathbf{V}\}$ is a subset of that generated by $\mathbf{X}$. By the same argument and using (32)

$$\mathbb{V}\text{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})$$

$$= \mathbb{E}(\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \mathbb{E}(h(\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})$$

$$= \int_{\text{supp}(f_\mathbf{X}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) \times f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1/t^{(0)}(\mathbf{V}, \mathbf{s}_0, f))$$

where $\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x})$. Using (32) again for the first term in (33) obtains formula (10) and (13).

To see that (7), (10), and (13) 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 (13)) which are continuous by assumption. In consequence, the parameter depending integrals (12) and (13) are well defined and continuous if (1) $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \cdot)$ is integrable for all $\mathbf{V} \in \mathcal{S}(p, q), \mathbf{s}_0 \in$ 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 [Heu95, 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 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 (13) 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 $\Sigma_\mathbf{x} > 0$. Then, $\tilde{L}(\mathbf{V}, \mathbf{s}_0, f)$ in (10) is continuous, which results in (8) also being well defined and continuous by virtue of it being a parameter depending integral following the same arguments as above. Moreover, (14) exists as the minimizer of a continuous function over the compact set $\mathcal{S}(p, q)$. Then, (8) can be writen as

$$L_\mathcal{F}^*(\mathbf{V}, f) = \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}} \left(\mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2 \right) + \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}} \left(\text{Var}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \right) \right) \tag{34}$$

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 (34) is constant with respect to $\mathbf{V}$. By the law of total variance,

$$\text{Var}(\tilde{\epsilon}) = \mathbb{E}\left(\text{Var}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \right) + \text{Var}\left(\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \right)$$
$$= \mathbb{E}\left(\text{Var}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) \right) \tag{35}$$

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 (35) into (34) obtains

$$L_\mathcal{F}^*(\mathbf{V}, f_t) = \mathbb{E}\left(\mu_2(\mathbf{V}, \mathbf{X}, f_t) - \mu_1(\mathbf{V}, \mathbf{X}, f_t)^2 \right) + \text{Var}(\tilde{\epsilon})$$
$$= \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}} \left(\text{Var}\left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \right) \right) + \text{Var}(\tilde{\epsilon}) \tag{36}$$

Next we show that (8), or, equivalently (36)), 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 (36)

$$\text{Var}\left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \right) = \text{Var}\left(g(\widetilde{\mathbf{B}}^T \mathbf{X}) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \right)$$
$$= \text{Var}\left(g(\widetilde{\mathbf{B}}^T \mathbf{s}_0 + \widetilde{\mathbf{B}}^T \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) \right) \geq 0 \tag{37}$$

If (37) 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 $L_\mathcal{F}^*(\mathbf{V}, f) = \text{Var}(\tilde{\epsilon})$. Since $\mathbf{s}_0$ is arbitrary yet constant, the same inequality holds for (8); that is, (8) attains its minimum for $\mathbf{V}$ such that $\mathbf{V}$ and $\widetilde{\mathbf{B}}^T$ are orthogonal. Since $\text{span}\{\widetilde{\mathbf{B}}^T\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$, (14) follows. $\qquad\square$

Proof of Theorem 6. Under assumptions (E.1), (E.2), and (E.3), (7) is well defined and continuous by arguments analogous to those in the proof of Theorem 4. Therefore (17) 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 $L_\mathcal{F}(\mathbf{Z}) = \min_{\mathbf{V} \in \mathcal{S}(p, p-k)} L_\mathcal{F}(\mathbf{V})$. By (14) and (15) in Theorem 4, $L_\mathcal{F}^*(\mathbf{V}, f_t)$, considered as a function from $\mathbb{R}^{p \times (p-k_t)}$, is minimized by an orthonormal base of $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}^\perp$ 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 4, we obtain that $L_\mathcal{F}^*(\mathbf{V}, f_t)$, 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 $\mathcal{F}$ characterises $\mathcal{S}_{Y|\mathbf{X}}$, 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,

$$\min_{\mathbf{V} \in \mathcal{S}(p, p-k)} L_\mathcal{F}(\mathbf{V}) = L_\mathcal{F}(\mathbf{Z}) = \mathbb{E}_{t \sim F_T}\left(L_\mathcal{F}^*(\mathbf{Z}, f_t) \right)$$

$$= \int_A \underbrace{L_\mathcal{F}^*(\mathbf{Z}, f_t)}_{>L_\mathcal{F}^*(\mathbf{U}, f_t)} dF_T(t) + \int_{A^c} \underbrace{L_\mathcal{F}^*(\mathbf{Z}, f_t)}_{=L_\mathcal{F}^*(\mathbf{U}, f_t)} dF_T(t) > \mathbb{E}_{t \sim F_T}\left(L_\mathcal{F}^*(\mathbf{U}, f_t) \right),$$

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

Next we introduce notation and auxiliary lemmas for the proof of Theorem 8. We suppose all assumptions of Theorem 8 hold. We generically use the letter "C" to denote constants.

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

$$\tilde{Y}_i = f(Y_i) = g(\widetilde{\mathbf{B}}^T \mathbf{X}_i) + \tilde{\epsilon}_i \tag{38}$$

with $\mathrm{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(\tilde{Y}|\mathbf{X})} = \mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}$. 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

$$t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = \frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^n K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) \tilde{Y}_i^l, \tag{39}$$

which is the sample version of (12) for $l = 0, 1, 2$. Eqn. (21) can be expressed as

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)}, \tag{40}$$

Lemma 12. *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 (31).

Proof of Lemma 12. By (31), $\|\mathbf{P}_{\mathbf{U}}(\mathbf{x} - \mathbf{s}_0)\|^2 = \|\mathbf{U}\mathbf{r}_2\|^2 = \|\mathbf{r}_2\|^2$. Further

$$\mathbb{E}\left(Z_n(\mathbf{V}, \mathbf{s}_0)\right) = \frac{1}{h_n^{(p-q)/2}} \int_{supp(f_{\mathbf{x}})} g(\mathbf{x})^l K(\|\mathbf{P}_{\mathbf{U}}(\mathbf{x} - \mathbf{s}_0)/h_n^{1/2}\|^2) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

$$= \frac{1}{h_n^{(p-q)/2}} \int_{supp(f_{\mathbf{x}}) \cap \mathbb{R}^{p-q}} \int_{supp(f_{\mathbf{x}}) \cap \mathbb{R}^q} g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)^l K(\|\mathbf{r}_2/h_n^{1/2}\|^2) \times$$
$$f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

$$= \int_{supp(f_{\mathbf{x}}) \cap \mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2) \int_{supp(f_{\mathbf{x}}) \cap \mathbb{R}^q} g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2)^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2$$

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. $\square$

Lemma 13. *Assume (E.1), (E.2), (E.3), (E.4), (H.1) and (K.1) hold. Then, there exists a constant $C > 0$, such that*

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

for $n > n^\star$ and $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$, $l = 0, 1, 2$, in (39).

Proof of Lemma 13. Since a continuous function attains a finite maximum over a compact set, $\sup_{\mathbf{x} \in supp(f_{\mathbf{x}})} |g(\widetilde{\mathbf{B}}^T \mathbf{x})| < \infty$. Therefore,

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

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

$$\mathbb{V}\mathrm{ar}\left(nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f)\right) = n\mathbb{V}\mathrm{ar}\left(\tilde{Y}^l K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \leq n\mathbb{E}\left(\tilde{Y}^{2l} K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right)$$

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

$$= n \sum_{u=0}^{2l} \binom{2l}{u} C^u \mathbb{E}\left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \tag{41}$$

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 12 with

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

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

$$\mathbb{E} \left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) K^2 \left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n} \right) \right) = h_n^{(p-q)/2} \mathbb{E}(Z_n(\mathbf{V}, \mathbf{s}_0))$$

$$= h_n^{(p-q)/2} \int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^{p-q}} K^2(\|\mathbf{r}_2\|^2) \times$$

$$\int_{\text{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} g_{2l-u}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h_n^{1/2}\mathbf{U}\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \tag{42}$$

$$\leq h_n^{(p-q)/2} C$$

since all integrands in (42) 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 (42) into (41) yields

$$\mathbb{V}\text{ar} \left(nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f) \right) \leq nh_n^{(p-q)/2} \underbrace{\sum_{u=0}^{2l} \binom{2l}{u} C^u C}_{=C} = nh_n^{(p-q)/2} C \tag{43}$$

$\square$

In Lemma 14 we show that $d_i(\mathbf{V}, \mathbf{s}_0)$ in (19) is Lipschitz in its inputs under assumption (E.2).

Lemma 14. *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 $\mathbf{s}_0, \mathbf{s}_j \in \text{supp}(f_{\mathbf{X}}) \subset \mathbb{R}^p$ 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 (19)

Proof of Lemma 14.

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq \left| \|\mathbf{X}_i - \mathbf{s}_0\|^2 - \|\mathbf{X}_i - \mathbf{s}_j\|^2 \right| +$$

$$\left| \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{P}_{\mathbf{V}}(\mathbf{X}_i - \mathbf{s}_0) \rangle - \langle \mathbf{X}_i - \mathbf{s}_j, \mathbf{P}_{\mathbf{V}_j}(\mathbf{X}_i - \mathbf{s}_j) \rangle \right| = I_1 + I_2 \tag{44}$$

where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^p$. We bound the first term on the right hand side of (44) 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).

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

$$\leq 2\|\mathbf{X}_i\|\|\mathbf{s}_0 - \mathbf{s}_j\| + 2C_1\|\mathbf{s}_0 - \mathbf{s}_j\| \leq 2C_1\delta + 2C_1\delta = 4C_1\delta$$

by Cauchy-Schwartz and the reverse triangular inequality for which $\left| \|\mathbf{s}_0\|^2 - \|\mathbf{s}_j\|^2 \right| = \left| \|\mathbf{s}_0\| - \|\mathbf{s}_j\| \right| (\|\mathbf{s}_0\| + \|\mathbf{s}_j\|) \leq \|\mathbf{s}_0 - \mathbf{s}_j\| 2C_1$. The second term in (44) satisfies

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

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

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

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

To show Theorems 8 and 15, we use the **Bernstein inequality** [S.N27]. 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,

$$P(|S_n - E_n| > t) < 2\exp\left(-\frac{t^2/2}{V_n + bt/3}\right) \tag{45}$$

Assumption (K.2) yields

$$|K(u) - K(u')| \leq K^*(u')\delta \tag{46}$$

for all u, u' with $|u - u'| < \delta \leq L_2$ and $K^*(\cdot)$ is a bounded and integrable kernel function [see [Han08]]. 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 15 and 16 we show that (39) converges uniformly in probability to (12) by showing that the variance and bias terms vanish uniformly in probability, respectively.

Lemma 15. *Under the assumptions of Theorem 8,*

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) \right| = O_P(a_n), \; l = 0, 1, 2 \tag{47}$$

Proof of Lemma 15. 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 (45). If the function f in (38) 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

$$t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) = (1/nh_n^{(p-q)/2}) \sum_i K(\|\mathbf{P_U}(\mathbf{X}_i - \mathbf{s}_0)\|^2/h_n)\tilde{Y}_i^l 1_{\{|\tilde{Y}_i|^l \leq \tau_n\}} \tag{48}$$

be the truncated version of (39) 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 (39). Therefore $R_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) \leq M_1 \tilde{R}_n^{(l)}$ due to (K.1) and

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) \right| \leq M_1(\tilde{R}_n^{(l)} + \mathbb{E}\tilde{R}_n^{(l)})$$
$$+ \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) \right| \tag{49}$$

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

$$\mathbb{E}\tilde{R}_n^{(l)} = \frac{1}{h_n^{(p-q)/2}}\mathbb{E}\left(|\tilde{Y}_i|^l 1_{\{|\tilde{Y}_i|^l > \tau_n\}}\right) \leq \frac{1}{h_n^{(p-q)/2}}\sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})}\sqrt{\mathbb{P}(|\tilde{Y}_i|^l > \tau_n)}$$

$$\leq \frac{1}{h_n^{(p-q)/2}}\sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})}\left(\frac{\mathbb{E}(|\tilde{Y}_i|^{4l})}{a_n^{-4}}\right)^{1/2} = o(a_n), \tag{50}$$

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 $\tilde{Y}_i = f(Y_i) \leq \sup_{f \in \mathcal{F}}|f(Y_i)| < C < \infty$.

Therefore, the first two terms of the right hand side of (49) converge to 0 with rate a_n by (50) and Markov's inequality. From this point on, $\tilde{Y}_i$ will denote the truncated version $\tilde{Y}_i 1_{\{|\tilde{Y}_i| \leq \tau_n\}}$ and we do not distinguish the truncated from the untruncated $t_n(\mathbf{V}, \mathbf{s}_0)$ since this truncation results in an error of magnitude a_n.

(ii) For the discretization step we cover the compact set $A = \mathcal{S}(p, q) \times \text{supp}(f_{\mathbf{X}})$ by finitely many balls, which is possible by (E.2) and the compactness of $\mathcal{S}(p, q)$. Let $\delta_n = a_n h_n$ and $A_j = \{\mathbf{V} : \|\mathbf{P_V} - \mathbf{P_{V_j}}\| \leq \delta_n\} \times \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}_j\| \leq \delta_n\}$ be a cover of A with ball centers $\mathbf{V}_j \times \mathbf{s}_j$. Then, $A \subset \bigcup_{j=1}^{N} A_j$ and the number of balls can be bounded by $N \leq C \delta_n^{-d}\delta_n^{-p}$ for some constant $C \in (0, \infty)$, where $d = \dim(\mathcal{S}(p, q)) = pq - q(q+1)/2$. Let $\mathbf{V} \times \mathbf{s}_0 \in A_j$. Then by Lemma 14 there exists $0 < C_2 < \infty$, such that

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2\delta_n \tag{51}$$

for d_i in (19). Under (K.2), which implies (46), inequality (51) yields

$$\left|K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) - K\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right)\right| \leq K^*\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right)C_2 a_n \tag{52}$$

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 (52) yields

$$|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)| \le C_2 a_n r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) \tag{53}$$

Since K^* fulfills (K.1) except for continuity, an analogous argument as in the proof of Lemma 12 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, (53) and integrability of r_n^l, we obtain

$$\left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \le \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right| + \left|\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right|$$

$$+ \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \le C_2 a_n \left(|r_n| + |\mathbb{E}(r_n)|\right) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right|$$

$$\le C_2 a_n(|r_n - \mathbb{E}(r_n)| + 2|\mathbb{E}(r_n)|) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right|$$

$$\le 2C_3 a_n + |r_n - \mathbb{E}(r_n)| + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \tag{54}$$

for any constant $C_3 > C_2 \mathbb{E}(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$ and n such that $C_2 a_n \le 1$, since $a_n^2 = o(1)$, which in turn yields that there exists $0 < C_3 < \infty$ such that (54) holds.

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

$$\mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n)$$

$$\le \sum_{j=1}^N \mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n)$$

$$\le N \max_{1 \le j \le N} \mathbb{P}(\sup_{\mathbf{V} \times \mathbf{s}_0 \in A_j} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)| > 3C_3 a_n) \tag{55}$$

$$\le N \left(\max_{1 \le j \le N} \mathbb{P}(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n) + \max_{1 \le j \le N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\right) \le$$

$$C \delta^{-(d+p)} \left(\max_{1 \le j \le N} \mathbb{P}(|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)| > C_3 a_n) + \max_{1 \le j \le N} \mathbb{P}(|r_n - \mathbb{E}(r_n)| > C_3 a_n)\right)$$

by the subadditivity of probability for the first inequality and (54) for the third inequality above, where the last inequality is due to $N \le C \delta_n^{-d} \delta_n^{-p}$ for a cover of A.

Finally, we bound the first and second term in the last line of (55) by the Bernstein inequality (45). For the first term in the last line of (55), 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 $|Z_i| \le b = M_1 \tau_n = M_1/a_n$ by (K.1) and the truncation step (i). For $V_n = \mathbb{V}\text{ar}(S_n)$, Lemma 13 yields $nh_n^{(p-q)/2} C \ge V_n$ with $C > 0$, and set $t = C_3 a_n nh_n^{(p-q)/2}$. The Bernstein inequality (45) yields

$$\mathbb{P}\left(\left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| > C_3 a_n\right) < 2\exp\left(\frac{-t^2/2}{V_n + bt/3}\right) \le$$

$$2\exp\left(-\frac{(1/2)C_3^2 a_n^2 n^2 h_n^{(p-q)}}{nh_n^{(p-q)/2}C + (1/3)M_1 \tau_n C_3 a_n nh_n^{(p-q)/2}}\right) \le 2\exp\left(-\frac{(1/2)C_3 \log(n)}{C/C_3 + M_1/3}\right) = 2n^{-\gamma(C_3)}$$

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

For the second term in the last line of (55), set $Z_i = Y_i^l K^*(d_i(\mathbf{V}_j, \mathbf{s}_j)/h_n)$ in (45) 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), $h_n^{(p-q)/4} \le 1$ for n large and (H.2) implies $1/(nh_n^{(p-q)/2}) \le 1$ for n large, therefore $h_n^{-1} \le n^{2/(p-q)} \le n^2$ since $p - q \ge 1$. Then, $\delta_n^{-1} = (a_n h_n)^{-1} \le n^{1/2} h_n^{-1} h_n^{(p-q)/4} \le n^{5/2}$. Therefore, (55) is smaller than $4 C \delta_n^{-(d+p)} n^{-\gamma(C_3)} \le 4Cn^{5(d+p)/2 - \gamma(C_3)}$. 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. $\qquad \square$

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 15 is no longer necessary.

Lemma 16. *Under (E.1), (E.2), (E.3), (E.4), (H.1), (K.1), and $\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)d\mathbf{r}_2 = 1$,*

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A} \left| t^{(l)}(\mathbf{V},\mathbf{s}_0) + 1_{\{l=2\}}\tilde{h}(\mathbf{V},\mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)\right| = O(h_n), \quad l = 0,1,2 \tag{56}$$

where $t^{(l)}(\mathbf{V},\mathbf{s}_0)$ and $\tilde{h}(\mathbf{V},\mathbf{s}_0)$ are defined in Theorem 4.

Proof of Lemma 16. 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 (31).

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

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

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

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

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

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

Then

$$\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right) = \int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\tilde{g}(\mathbf{r}_1,h_n^{1/2}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 \tag{57}$$

holds by Lemma 12 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 (57) 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), $|\mathbf{r}_2^T\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{r}_1,\xi)\mathbf{r}_2| \leq C\|\mathbf{r}_2\|^2$ for $C = \sup_{\mathbf{x},\mathbf{y}}\|\nabla_{\mathbf{r}_2}^2\tilde{g}(\mathbf{x},\mathbf{y})\| < \infty$, since a continuous function over a compact set is bounded. Then, $R(\mathbf{V},\mathbf{s}_0) \leq CC_4\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\|\mathbf{r}_2\|^2 d\mathbf{r}_2 < \infty$ for some $C_4 > 0$, since the integral over $\mathbf{r}_1$ is over a compact set by (E.2). $\square$

Lemma 17 follows directly from Lemmas 15 and 16 and the triangle inequality.

Lemma 17. *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 18. *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*

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

23

and

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| \tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{s}_0) - \tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{s}_0) \right| = o_P(1) \tag{58}$$

where $\bar{y}_l(\mathbf{V}, \mathbf{s}_0)$, $\mu_l(\mathbf{V}, \mathbf{s}_0)$, $\tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{s}_0)$ *and* $\tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{s}_0)$ *are defined in* (21), (11), (22) *and* (10), *respectively.*

Proof of Theorem 18. 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 (12), 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,

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)} = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)} \tag{59}$$

We consider the numerator and enumerator of (59) separately. By Lemma 17

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

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

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 (59), we obtain

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

For $l = 2$, $\tilde{Y}_i^2 = g(\widetilde{\mathbf{B}}^T \mathbf{X}_i)^2 + 2g(\widetilde{\mathbf{B}}^T \mathbf{X}_i)\tilde{\epsilon}_i + \tilde{\epsilon}_i^2$, and (58) follows from (10). $\qquad\square$

Lemma 19. *Under (E.1), (E.2), (E.4), there exists* $0 < C_5 < \infty$ *such that*

$$\left| \mu_l(\mathbf{V}, \mathbf{s}_0) - \mu_l(\mathbf{V}_j, \mathbf{s}_0) \right| \leq C_5 \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{60}$$

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

Proof. From the representation $\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)$ in (16) instead of $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$, we consider $\mu_l(\mathbf{V}, \mathbf{s}_0) = \mu_l(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)$ as a function on the Grassmann manifold since $\mathbf{P}_{\mathbf{V}} \in Gr(p, q)$. Then,

$$\left| \mu_l(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0) - \mu_l(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0) \right| = \left| \frac{\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)} - \frac{\tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)}{\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0)} \right|$$

$$\leq \frac{\sup |\tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)|}{(\inf \tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0))^2} \left| \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0) - \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0) \right|$$

$$+ \frac{\sup \tilde{t}^{(l)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0)}{(\inf \tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0))^2} \left| \tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}}, \mathbf{s}_0) - \tilde{t}^{(0)}(\mathbf{P}_{\mathbf{V}_j}, \mathbf{s}_0) \right| \tag{61}$$

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

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

$$\leq L \int_{\text{supp}(f_{\mathbf{X}})} \|(\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j})\mathbf{r}\| d\mathbf{r} \leq L \left(\int_{\text{supp}(f_{\mathbf{X}})} \|\mathbf{r}\| d\mathbf{r} \right) \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \tag{62}$$

where the last inequality is due to the sub-multiplicativity of the Frobenius norm and the integral being finite by (E.2). Plugging (62) in (61) and collecting all constants into C_5 yields (60). $\qquad\square$

Proof of Theorem 8. By (23) and (7),

$$|L_n^*(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}, f)| \leq \left| \frac{1}{n} \sum_i \left(\tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f) \right) \right|$$
$$+ \left| \frac{1}{n} \sum_i \left(\tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f) - \mathbb{E}(\tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{X}, f)) \right) \right| \tag{63}$$

By Theorem 18,

$$\left| \frac{1}{n} \sum_i \tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f) \right| \leq \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| \tilde{L}_{n,\mathcal{F}}(\mathbf{V}, \mathbf{s}_0, f) - \tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{s}_0, f) \right| = o_P(1) \tag{64}$$

The second term in (63) 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 15 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 15. By Lemma 19,

$$|\mu_l(\mathbf{V}, \mathbf{X}_i) - \mu_l(\mathbf{V}_j, \mathbf{X}_i)| \leq C_5 \|\mathbf{P}_\mathbf{V} - \mathbf{P}_{\mathbf{V}_j}\| \tag{65}$$

Let $G_n(\mathbf{V}, f) = \sum_i \tilde{L}_{\mathcal{F}}(\mathbf{V}, \mathbf{X}_i, f)/n$ with $\mathbb{E}(G_n(V)) = L_{\mathcal{F}}^*(\mathbf{V}, f)$. Using (65) and following the same steps as in the proof of Lemma 15 we obtain

$$|G_n(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}, f)| \leq |G_n(\mathbf{V}, f) - G_n(\mathbf{V}_j, f)|$$
$$+ |G_n(\mathbf{V}_j, f) - L_{\mathcal{F}}^*(\mathbf{V}_j, f)| + |L_{\mathcal{F}}^*(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}_j, f)|$$
$$\leq 2C_6\tilde{a}_n + |G_n(\mathbf{V}_j, f) - L_{\mathcal{F}}^*(\mathbf{V}_j, f)| \tag{66}$$

for $\mathbf{V} \in B_j$ and some $C_6 > C_5$. Inequality (66) leads to

$$\mathbb{P}\left(\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |G_n(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}, f)| > 3C_6\tilde{a}_n \right) \leq C\,N\,\mathbb{P}(\sup_{\mathbf{V} \in B_j} |G_n(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}, f)| > 3C_6\tilde{a}_n)$$
$$\leq C\,n^{d/2}\mathbb{P}(|G_n(\mathbf{V}_j, f) - L_{\mathcal{F}}^*(\mathbf{V}_j, f)| > C_6\tilde{a}_n)$$
$$\leq C\,n^{d/2}n^{-\gamma(C_6)} \to 0 \tag{67}$$

where the last inequality in (67) is due to (45) with $Z_i = \tilde{L}_{\mathcal{F}}(\mathbf{V}_j, \mathbf{X}_i, f)$, which is bounded since $\tilde{L}_{\mathcal{F}}(\cdot, \cdot, f)$ 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, $\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f) - L_{\mathcal{F}}^*(\mathbf{V}, f)| \leq o_P(1) + O_P(\tilde{a}_n)$ which implies Theorem 8. □

Proof of Theorem 9. We apply [Ame85, 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., $\mathbf{P}_{\text{span}\{\hat{\mathbf{B}}_{k_t}^t\}^\perp} = \operatorname{argmin} L_n^*(\mathbf{P}_\mathbf{V}, f) \to \mathbf{P}_{\text{span}\{\mathbf{B}\}^\perp} = \operatorname{argmin} L^*(\mathbf{P}_\mathbf{V}, f_t)$ 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 $(Y_i, \mathbf{X}_i^T)_{i=1,\dots,n}$, 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 $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}^\perp$.

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 (16). 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 8 obtains the uniform convergence and Theorem 4 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. $\|\mathbf{P}_{\text{span}\{\hat{\mathbf{B}}_{k_t}^t\}} - \mathbf{P}_{\text{span}\{\widetilde{\mathbf{B}}\}}\| = \|\hat{\mathbf{B}}_{k_t}^t(\hat{\mathbf{B}}_{k_t}^t)^T - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T\| = \|(\mathbf{I}_p - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T) - (\mathbf{I}_p - \hat{\mathbf{B}}_{k_t}^t(\hat{\mathbf{B}}_{k_t}^t)^T)\| = \|\mathbf{P}_{\text{span}\{\widetilde{\mathbf{B}}\}^\perp} - \mathbf{P}_{\text{span}\{\hat{\mathbf{B}}_{k_t}^t\}^\perp}\|$). Thus, all three conditions are met and the result is obtained. □

Proof of Theorem 10. Let $(t_j)_{j=1,\ldots,m_n}$ be an i.i.d. sample from F_T and write

$$|L_{n,\mathcal{F}}(\mathbf{V}) - L_{\mathcal{F}}(\mathbf{V})| = \left| \frac{1}{m_n} \sum_{j=1}^{m_n} \left(L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j}) \right) \right|$$
$$+ \left| \frac{1}{m_n} \sum_{j=1}^{m_n} \left(L^*(\mathbf{V}, f_{t_j}) - \mathbb{E}_{t \sim F_T}(L^*(\mathbf{V}, f_t)) \right) \right| \tag{68}$$

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 assumption $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$, and the triangle inequality. 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 (68) and let $\delta > 0$. By Markov's and triangle inequalities and Fatou's Lemma,

$$\limsup_n \mathbb{P} \left(\sup_{\mathbf{V} \in \mathcal{S}(p,q)} \left| \frac{1}{m_n} \sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j}) \right| > \delta \right)$$

$$\leq \frac{1}{\delta} \limsup_n \mathbb{E}_{F_T} \left(\mathbb{E}(\sup_{\mathbf{V} \in \mathcal{S}(p,q)} \left| \frac{1}{m_n} \sum_{j=1}^{m_n} L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j}) \right|) \right) \text{ :Markov inequality}$$

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

$$= \frac{1}{\delta} \limsup_n \mathbb{E}_{F_T} \left(\mathbb{E}(\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})|) \right)$$

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

since by Theorem 8 it holds $\limsup_n \sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j})| = 0$.

For the second term on the right hand side of (68) we apply Theorem 2 of [Jen69] in [MMW$^+$63, p. 40]:

Theorem 20. *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 4. If $L^*(\mathbf{V}, f_{t_j}) \leq h(t_j)$, where $h(t_j)$ is integrable with respect to F_T, then*

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

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 (68), $Z_j(\mathbf{V}) = L^*(\mathbf{V}, f_{t_j}) < 4M^2$. 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 4 and the parameter space $\mathcal{S}(p,q)$ is compact. Then by Theorem 20,

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

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 $n \to \infty$. $\square$

Proof of Theorem 11. The proof is directly analogous to the proof of Theorem 9. The uniform convergence of the target function $L_{n,\mathcal{F}}(\mathbf{V})$ is obtained by Theorem 10. The minimizer over $Gr(p,q)$ and its uniqueness derive from Theorem 6. $\qquad\square$

Proof of Theorem 7. In this proof we supress the dependence on f in the notation. The Gaussian kernel K satisfies $\partial_z K(z) = -zK(z)$. From (20) and (22) we have $\tilde{L}_{n,\mathcal{F}} = \bar{y}_2 - \bar{y}_1^2$ 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,

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

$$= -\frac{1}{h_n^2} \left(\sum_i \tilde{Y}_i^l w_i d_i \nabla d_i - \sum_j \tilde{Y}_j^l w_j \sum_i w_i d_i \nabla d_i \right) = -\frac{1}{h_n^2} \sum_i (\tilde{Y}_i^l - \bar{y}_l) w_i d_i \nabla d_i \qquad (69)$$

Then, $\nabla \tilde{L}_n = \nabla \bar{y}_2 - 2\bar{y}_1 \nabla \bar{y}_1$, and inserting $\nabla \bar{y}_l$ from (69) yields $\nabla \tilde{L}_n = (-1/h_n^2) \sum_i (Y_i^2 - \bar{y}_2 - 2\bar{y}_1(Y_i - \bar{y}_1)) w_i d_i \nabla d_i = (1/h_n^2) \left(\sum_i \left(\tilde{L}_n - (Y_i - \bar{y}_1)^2 \right) w_i d_i \nabla d_i \right)$, since $Y_i^2 - \bar{y}_2 - 2\bar{y}_1(Y_i - \bar{y}_1) = (Y_i - \bar{y}_1)^2 - \tilde{L}_n$. $\qquad\square$