TECHNISCHE UNIVERSITÄT WIEN

DISSERTATION

Sufficient Dimension Reduction using Conditional Variance Estimation and related concepts

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften unter der Leitung von

Univ. Prof. Ph.D Efstathia Bura
E101 – Institut für Angewandte Statistik, TU Wien

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Lukas Fertl
Matrikelnummer: 1151896
[redacted]

Diese Dissertation haben begutachtet:

1. **Univ. Prof., Ph.D. Efstathia Bura**
   Institut für Stochastik und Wirtschaftsmathematik, Technische Universität Wien
2. **Univ. Prof., Ph.D. Liliana Forzani**
   Departamento de matematica, Facultad de Ingenieria Quimica, Universidad Nacional del Litoral
3. **Univ. Prof., Ph.D. Bing Li**
   Institut of Statistics, Pennsylvania State University

Wien, June 15, 2021

Kurzfassung

In der Regression untersucht man die bedingte Verteilung der Zielvariable gegeben den Prädiktoren, um z.B. Prognosen zu erhalten. Regression ist einer der meist studierten und angewandten Gebiet der Statistik. Die Modellierung von hochdimensionalen Daten, insbesonders bei einem nichtlinearern Zusammenhang, ist herausfordernd falls die Anzahl der Prädiktoren (p) groß ist. Suffiziente Dimensionsreduktion (SDR) ersetzt den hochdimensionalen Prädiktorvektor durch eine niedrigdimensionalere Projektion, ohne Information über die Zielvariable zu verlieren.

Diese Arbeit entwickelt neue SDR Ansätze, den *conditional variance* und *ensemble conditional variance* estimator, für die Identifikation und Schätzung der linearen suffizienten Reduktion sowohl für den bedingten Erwartungswert alsauch die bedingte Verteilngsfunktion der Zielvariable gegeben den hochdimensionalen Prädiktoren. Für beide Schätzer wird die Konsistenz bewiesen. Weiters, wird eine neuer Schätzer, der eine Kombination aus suffizienter Dimensionsreduktion und Neuronalen Netzten ist, vorgestellt. Alle drei Schätzer sind kompetitive im Vergleich zu momentanen state-of-the-art SDR Schätzern.

Abstract

Regression concerns modeling the conditional distribution of a target variable, the response, given a set of other variables, the predictors. Regression is the most widely used approach in Statistical applications. As such, it has been extensively studied since the field of Statistics came to existence. Modeling high-dimensional data is challenging, especially when they are nonlinearly related. Sufficient dimension reduction (SDR) considers regressions where the number of predictors (p) is large and replaces the high dimensional predictor by a lower dimensional reduction (function) without loss of information for the response.

This thesis develops novel SDR approaches, the *conditional variance* and *ensemble conditional variance* estimators, for the identification and estimation of linear sufficient reductions both for the conditional mean and the conditional cumulative distribution function of the response given the multidimensional predictors. The consistency of both estimators is shown. Moreover, a combination of sufficient dimension reduction with neural networks is derived, which leverages the advantages of both in order to predict the response in the presence of abundant predictors and observations. All three proposed estimators are competitive with respect to current state-of-the-art methods in SDR methodology.

Acknowledgement

First, I would like to thank my family, especially my mother Elisabeth Fertl, for the support throughout my studies. Moreover, I would like to thank my supervisor Efstathia Bura and colleagues for the support during my PhD studies. Further, I gratefully acknowledge the support of the Austrian Science Fund (FWF P 30690-N35) and thank Daniel Kapla, my colleague and friend, for his cowork and programming assistance. Daniel Kapla also co-authored the CVarE `R` package that implements the conditional variance estimation method.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am 15. Juni 2021

Lukas Fertl

Contents

1 Introduction

In this chapter, the historic development of SDR are presented. We start by introducing the concepts of the central and mean subspaces. The former are dimension reduction subspaces that preserve all the information, whereas the latter is a subset of the former that captures only the information contained in the conditional mean of the response. Most SDR methods are inverse regression based, i.e. regressing the predictors on the response, that require assumptions on the marginal distribution of the predictors or assume knowledge of the family of distributions for the response and the predictors, and semiparametric, such as mean average variance estimation (MAVE) [XTLZ02], which are based on the forward regression model of the response on the predictors. The former operate under restrictive assumptions and are usually computationally easy to implement. The latter are computationally more expensive but frequently lead to better estimates of the reduction with the additional advantage of automatically predicting the response. They also are few in number, with MAVE remaining the gold standard so far. This thesis contributes three new forward based methods: the *conditional variance* in chapter 2, *neural net-sufficient dimension reduction* in chapter 3 and *ensemble conditional variance* estimation in chapter 4.

We start by defining the notation that is used throughout the thesis in Section 1.1. In Section 1.2, we review regression problems and the motivation for sufficient dimension reduction is presented. In Section 1.3 sufficient dimension reduction is defined and the special case of *linear* sufficient dimension reduction is introduced and discussed in Section 1.4, which leads to the concept of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. In Section 1.5 the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ is defined and in Section 1.6 a historic account of sufficient dimension reduction are presented. In the following sections some important sufficient dimension reduction estimation methods are summarized. Finally, in Section 1.16, the contributions of this thesis are outlined.

1.1 Notation and Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, throughout we suppose $(Y, \mathbf{X}^T)^T$ has a joint distribution, where $Y : \Omega \to \mathbb{R}$ denotes a univariate response and $\mathbf{X} : \Omega \to \mathbb{R}^p$ a p-dimensional covariate vector with a continuous distribution. We will denote the probability density function of $\mathbf{X}$ by $f_{\mathbf{X}}$ (if it exists) and denote its support by $\text{supp}(f_{\mathbf{X}})$. Moreover for random variables $\mathbf{X}, Y$ stochastic independence will be denoted as $Y \perp\!\!\!\perp \mathbf{X}$. The space $L^m(\Omega)$ is the set of all random variables on the probability space with finite m-th moment. Throughout $F(\cdot|\cdot)$ is the cumulative distribution function (cdf) of the first argument given the second, $\|\cdot\|$ denotes the Frobenius norm for matrices, Euclidean norm for vectors, scalar product refers to the euclidean scalar product, and $\perp$ denotes orthogonality with respect to the euclidean scalar product. Further $\mathbf{e}_j \in \mathbb{R}^p$ denotes the j-th standard basis vector with zeroes except on the j-th position, $\boldsymbol{\iota}_p = (1, 1, \ldots, 1)^T \in \mathbb{R}^p$, and $\mathbf{I}_p = (\mathbf{e}_1, \ldots, \mathbf{e}_p) \in \mathbb{R}^{p \times p}$ is

the p-dimensional identity matrix. For any full rank matrix $\mathbf{M}$, or linear subspace $\mathbf{M}$, we denote by $\mathbf{P}_\mathbf{M}$ the projection matrix on the column space of the matrix or on the subspace, i.e. $\mathbf{P}_\mathbf{M} = \mathbf{M}(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T \in \mathbb{R}^{p\times p}$ for $\mathbf{M} \in \mathbb{R}^{p\times q}$ with $\text{rank}(\mathbf{M}) = q$ [1].

For $q \leq p$, let

$$\mathcal{S}(p,q) = \{\mathbf{V} \in \mathbb{R}^{p\times q} : \mathbf{V}^T\mathbf{V} = \mathbf{I}_q\}, \tag{1.1}$$

denote the Stiefel manifold, that comprizes of all $p \times q$ matrices with orthonormal columns. $\mathcal{S}(p,q)$ is compact and $\dim(\mathcal{S}(p,q)) = pq - q(q+1)/2 = p(q-(q+1)/2)$ [see [Boo02] and Section 2.1 of [Tag11]]. Next the Grassmann manifold, used in the section 4.4.1, is defined and some properties are presented. Let

$$Gr(p,k) = \mathcal{S}(p,k)/\mathcal{S}(k,k) \tag{1.2}$$

denote the Grassmann manifold, i.e. all k-dimensional subspaces in $\mathbb{R}^p$. The Grassmann manifold is compact as a quotient space of the compact Stiefle manifold (see (1.1)) and the dimension is given by $\dim(Gr(p,k)) = \dim(\mathcal{S}(p,k)) - \dim(\mathcal{S}(k,k)) = k(p-(k+1)/2) - k(k-(k+1)/2) = k(p-k)$, for further information see [GH94]. We can identify a subspace $\mathbf{M} \in Gr(p,k)$ with the orthogonal projection matrix $\mathbf{P}_\mathbf{M}$ onto this subspace and for an orthonormal basis $\mathbf{B} \in \mathcal{S}(p,k)$ of $\mathbf{M}$ we have

$$\mathbf{P}_\mathbf{M} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T = \mathbf{B}\mathbf{B}^T = \mathbf{B}\underbrace{\mathbf{O}\mathbf{O}^T}_{=\mathbf{I}_k}\mathbf{B}^T = (\mathbf{B}\mathbf{O})(\mathbf{B}\mathbf{O})^T$$

for all orthogonal matrices $\mathbf{O} \in \mathcal{S}(k,k)$.

For any $\mathbf{V} \in \mathcal{S}(p,q)$, defined in (1.1), we generically denote a basis of the orthogonal complement of its column space $\text{span}\{\mathbf{V}\}$, by $\mathbf{U}$. That is, $\mathbf{U} \in \mathcal{S}(p,p-q)$ such that $\text{span}\{\mathbf{V}\} \perp \text{span}\{\mathbf{U}\}$ and $(\mathbf{V},\mathbf{U})$ an orthonormal base of $\mathbb{R}^p$, i.e. $\mathbf{U}^T\mathbf{V} = \mathbf{0} \in \mathbb{R}^{(p-q)\times q}, \mathbf{U}^T\mathbf{U} = \mathbf{I}_{p-q}$. For any $\mathbf{x}, \mathbf{s}_0 \in \mathbb{R}^p$ we can always write

$$\mathbf{x} = \mathbf{s}_0 + \mathbf{P}_\mathbf{V}(\mathbf{x}-\mathbf{s}_0) + \mathbf{P}_\mathbf{U}(\mathbf{x}-\mathbf{s}_0) = \mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2 \tag{1.3}$$

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

Further, we use the convention that $\mathbf{s}_0 \in \text{supp}(f_\mathbf{X})$ is understood to be as $\mathbf{s}_0$ is in the interior of the support since $\mathbf{s}_0$ acts as a placeholder for a datapoint $\mathbf{X}_i$ which takes values on the boundary of the support with probability 0.

Next we define a generalized eigenvalue problem.

Definition 1. *A generalized eigenvalue problem of two symmetric and positive definite matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathbb{R}^{p\times p}$ is defined as finding eigenvectors $\mathbf{v}_j \in \mathbb{R}^p$ and the corresponding eigenvalues $\lambda_j \in \mathbb{R}$ such that*

$$\begin{aligned} \mathbf{M}_1\mathbf{v}_j &= \lambda_j\mathbf{M}_2\mathbf{v}_j \quad \textit{for} \quad j = 1,\dots,p \\ \mathbf{M}_1\mathbf{V} &= \mathbf{M}_2\mathbf{V}\mathbf{\Lambda} \\ \textit{subject to} \quad \mathbf{V}^T\mathbf{M}_2\mathbf{V} &= \mathbf{I}_p \end{aligned} \tag{1.4}$$

where (1.4) is the matrix notation with the eigenvectors collected in $\mathbf{V} = (\mathbf{v}_1,\dots,\mathbf{v}_p) \in \mathbb{R}^{p\times p}$ and the eigenvalues in $\mathbf{\Lambda} = diag(\lambda_1,\dots,\lambda_p)$, see also [GKC19].

[1] If the matrix $\mathbf{M}$ is not of full rank then the term $(\mathbf{M}^T\mathbf{M})^{-1}$ is understood as the generalized inverse.

Remark. *A generalized eigenvalue problem (GEV) given by* $(\mathbf{M}_1, \mathbf{M}_2)$ *can be reformulated as an eigenvalue problem* $(\widetilde{\mathbf{M}}_1, \mathbf{I}_p)$

$$\begin{aligned}\mathbf{M}_1\mathbf{V} = \underbrace{\mathbf{M}_2}_{(\mathbf{M}_2^{1/2})^2}\mathbf{V}\mathbf{\Lambda} \quad &\iff \quad \underbrace{\mathbf{M}_2^{-1/2}\mathbf{M}_1\mathbf{M}_2^{-1/2}}_{\widetilde{\mathbf{M}}_1}\underbrace{\mathbf{M}_2^{1/2}\mathbf{V}}_{=\widetilde{\mathbf{V}}} = \underbrace{\mathbf{M}_2^{1/2}\mathbf{V}}_{=\widetilde{\mathbf{V}}}\Lambda \\ &\iff \widetilde{\mathbf{M}}_1\widetilde{\mathbf{V}} = \widetilde{\mathbf{V}}\Lambda \end{aligned} \tag{1.5}$$

with $\widetilde{\mathbf{M}}_1 = \mathbf{M}_2^{-1/2}\mathbf{M}_1\mathbf{M}_2^{-1/2}$ *and* $\widetilde{\mathbf{V}} = \mathbf{M}_2^{1/2}\mathbf{V}$. *Therefore given the solution* $\widetilde{\mathbf{V}}$ *of the eigenvalue problem* $(\widetilde{\mathbf{M}}_1, \mathbf{I}_p)$ *in* (1.5) *we can calculate the solution of* (1.4) *by* $\mathbf{V} = \mathbf{M}_2^{-1/2}\widetilde{\mathbf{V}}$. *Moreover* $\widetilde{\mathbf{V}} \in \mathcal{S}(p,p)$ *since it is the solution of an eigenvalue problem and we can deduce* $\mathbf{I}_p = \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}} = \mathbf{V}^T\mathbf{M}_2\mathbf{V}$, *i.e. the solution* $\mathbf{V}$ *of* (1.4) *is orthogonal with respect to the scalar product induced by* $\mathbf{M}_2$.

Remark. *A generalized eigenvalue problem (GEV) given by* $(\mathbf{M}_1, \mathbf{M}_2)$ *can be rewritten in terms of the Rayleigh quotient*

$$\operatorname{argmax} R(\mathbf{v}) = \operatorname{argmax} \frac{\mathbf{v}^T\mathbf{M}_1\mathbf{v}}{\mathbf{v}^T\mathbf{M}_2\mathbf{v}} \iff \mathbf{M}_1\mathbf{v} = \lambda\mathbf{M}_2\mathbf{v} \tag{1.6}$$

i.e. maximizing the quotient is equivalent to finding the eigenvector corresponding to the maximal eigenvalue of the GEV (the eigenvector corresponding to the smallest eigenvalue minimizes the quotient). This can be seen by taking the derivative of $R(\mathbf{v})$ *and setting it to* $\mathbf{0}$, *i.e.*

$$\begin{aligned}\nabla_{\mathbf{v}} R(\mathbf{v}) = 2\frac{\mathbf{M}_1\mathbf{v}(\mathbf{v}^T\mathbf{M}_2\mathbf{v}) - (\mathbf{v}^T\mathbf{M}_1\mathbf{v})\mathbf{M}_2\mathbf{v}}{(\mathbf{v}^T\mathbf{M}_2\mathbf{v})^2} = \mathbf{0} \quad &\iff \\ \mathbf{M}_1\mathbf{v}(\mathbf{v}^T\mathbf{M}_2\mathbf{v}) - (\mathbf{v}^T\mathbf{M}_1\mathbf{v})\mathbf{M}_2\mathbf{v} = \mathbf{0} \quad &\iff \\ \mathbf{M}_1\mathbf{v} = \underbrace{\frac{\mathbf{v}^T\mathbf{M}_1\mathbf{v}}{\mathbf{v}^T\mathbf{M}_2\mathbf{v}}}_{=\lambda}\mathbf{M}_2\mathbf{v} = \lambda\mathbf{M}_2\mathbf{v}\end{aligned}$$

1.2 Regression and dimension reduction

In this section a brief introduction to regression problems is given. Let $(\mathbf{Y}^T, \mathbf{X}^T)^T \in \mathbb{R}^{d+p}$ be a random vector with a joint continuous distribution where $\mathbf{Y} \in \mathbb{R}^d$ is called the response and $\mathbf{X} \in \mathbb{R}^p$ is the predictor vector.

Regression, in the broadest sense, is then the inference about the conditional distribution of the response $\mathbf{Y}$ given the predictors $\mathbf{X}$, i.e. we are interested in the distribution of $\mathbf{Y} \mid \mathbf{X}$. This is called the *forward regression* and $\mathbf{X}|\mathbf{Y}$ is the *inverse regression*. The conditional distribution captures all the information about $\mathbf{Y}$ given $\mathbf{X}$.

If we are interested in point prediction of $\mathbf{Y}$ given a specific value of $\mathbf{X} = \mathbf{x}$, measuring the quality of prediction by the mean squared error, the optimal prediction is given by the conditional mean $\mathbb{E}(\mathbf{Y} \mid \mathbf{X})$. By assuming $(\mathbf{Y}^T, \mathbf{X}^T)^T \in L^2(\Omega)$ it follows that the first

conditional moment exists [Kus14], and we model the response $\mathbf{Y} = (Y_1, \ldots, Y_d)$ via the predictors $\mathbf{X} = (X_1, \ldots, X_p)$, i.e.

$$\mathbf{Y} = \mathbb{E}(\mathbf{Y} \mid \mathbf{X}) + \underbrace{(\mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \mathbf{X}))}_{=\tilde{\epsilon}} = \mathbb{E}(\mathbf{Y} \mid \mathbf{X}) + \tilde{\epsilon} = g(\mathbf{X}) + \tilde{\epsilon} \tag{1.7}$$

where $\mathbb{E}(\mathbf{Y} \mid \mathbf{X}) = g(\mathbf{X}) = (g_1(\mathbf{X}), \ldots, g_d(\mathbf{X}))^T \in \mathbb{R}^d$, for measurable functions $g_j : \mathbb{R}^p \to \mathbb{R}$. $\tilde{\epsilon} = \mathbf{Y} - \mathbb{E}(\mathbf{Y} \mid \mathbf{X}) \in \mathbb{R}^d$ is called the residual. By the tower property of the conditional expectation the residual is conditionally centered, i.e. $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and centered $\mathbb{E}(\tilde{\epsilon}) = \mathbb{E}(\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X})) = \mathbf{0}$.

All quantities in (1.7) have finite variance and the residual is uncorrelated to $\mathbb{E}(Y \mid \mathbf{X}) = g(\mathbf{X})$, i.e. $\mathbb{E}(\tilde{\epsilon} g(\mathbf{X})^T) = \mathbb{E}(\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) g(\mathbf{X})^T) = \mathbf{0} \in \mathbb{R}^{d \times d}$. In general the components of the residual are not independent, if they are independent then (1.7) describes d independent regression problems.

In this work we assume from now on that the response is scalar, i.e. $d = 1$ and $\mathbf{Y} = Y_1 = Y$. The error term $\tilde{\epsilon}$ in (1.7) may still depend on $\mathbf{X}$ (see for example model (1.28) where the conditional variance of the residual, i.e. $\mathbb{V}\mathrm{ar}(Y - \mathbb{E}(Y \mid \mathbf{X}) \mid \mathbf{X})$ is a function of $\mathbf{X}$) so in general we do not have $\mathbf{X} \perp\!\!\!\perp \tilde{\epsilon}$. If $\mathbf{X} \perp\!\!\!\perp \tilde{\epsilon}$, we are in the classic setting with homoskedastic errors since the conditional variance equals the unconditional one and is therefore constant, i.e. $\mathbf{X} \perp\!\!\!\perp \tilde{\epsilon}$ implies $\mathbb{V}\mathrm{ar}(Y \mid \mathbf{X}) = \mathbb{V}\mathrm{ar}(Y)$. Further if $\mathbf{X}$ is independent of the residuals then $\mathbf{X}$ enters Y only through the first conditional moment $\mathbb{E}(Y \mid \mathbf{X})$, see Sections 1.4 and 1.5.

The conditional expectation $\mathbb{E}(Y \mid \mathbf{X}) = g(\mathbf{X})$ is called the regression or link function. In applications we are often interested in estimating g from an independent and identically distributed sample $(Y_i, \mathbf{X}_i)_{i=1}^n$ from model (1.7). In general, this is not a feasible task without further assumptions on g, or simplification through parametrization.

The former leads to a non-parametric problem where the search space is infinite dimensional, e.g. $C^2(\mathbb{R}^p)$ if we assume g is twice continuously differentiable.

The most prominent example of the latter approach is the *linear regression model* with homoskedastic error, which sets $g(\mathbf{x}) = \alpha + \boldsymbol{\beta}^T \mathbf{X}$, with $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. This modeling assumption drastically reduces the complexity of the problem since the search space is reduced to the finite dimensional space $\mathbb{R}^{p+1}$ from an infinite dimensional space.

Furthermore, if the dimension p of the covariate vector $\mathbf{X}$ is large, and we only assume only smoothness or differentiability of the link function g, we face the curse of dimensionality in nonparametric regression. A high dimensional input space, like $\mathbb{R}^p$ or $[0,1]^p$, is sparsely populated by the samples $(Y_i, \mathbf{X}_i)$ if p is large. This sparsity causes problems in estimating g if a parametric model is not assumed. If only smoothness is assumed, we can draw inference only from points nearby. Informally, in order to estimate $g(\mathbf{x})$ only points $\mathbf{z}$ with distance less than $1/m$ to $\mathbf{x}$, i.e. $\mathbf{z} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{z}\| \leq 1/m$, are relevant. Even when the sample size is very large, for common p values, the local neighborhoods about a point will contain very few, if any, observations with high probability. In mathematical terms this can be seen by considering the p dimensional hypercube $[0,1]^p$. If we want to place evenly spaced points in the hypercube the number of points required increases exponentially in p, i.e. to place evenly spaced points in $[0,1]^p$ with distance $1/m$ to each other requires

$(m+1)^p$ points. Therefore, the estimation of the link function g becomes intractable for large p.

A possible solution to the curse of dimensionality is dimension reduction, where we replace $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ by a lower dimensional input $R(\mathbf{X}) \in \mathbb{R}^k$ with $R : \mathbb{R}^p \to \mathbb{R}^k$ and $k << p$, prior to estimating the forward regression link function. If the reduction is such that we do not lose relevant information, we call the reduction a *sufficient* reduction and we can circumvent the curse of dimensionality. That is, we regress Y on the low dimensional input $R(\mathbf{X})$. In general the reduction function $R(\cdot)$ can be linear or non-linear. Assume for example that $Y = \sin(\|\mathbf{X}\|) + \epsilon$ with $\mathbf{X} \in \mathbb{R}^p$, then the reduction $R(\mathbf{X}) = \|\mathbf{X}\| \in \mathbb{R}$ would reduce the estimation problem to a one dimensional non-parametric regression for any p without loss of information. An example for a linear reduction is the model given by $Y = \exp(X_1) + \epsilon$, where the linear reduction is given by $R(\mathbf{X}) = \mathbf{e}_1^T \mathbf{X} = X_1$.

Moreover if the dimension of the covariate vector is reduced to $k = 1, 2$ then it is also possible to visualize the data by a scatter plot. This may be quite helpful in determining the appropriate techniques or methods for estimating the forward regression. Therefore dimension reduction is an important tool for high dimensional, especially non-parametric, regression problems in (1.7).

Nevertheless there are caveats, in real problems where only a sample $(Y_i, \mathbf{X}_i)_{i=1}^n$ is given and we do not know if there is a useful reduction function $R(\cdot)$, how it looks like, and the dimension k of the reduced input is also unknown. The first problem can only be solved by assuming that there exists a reduction function whereas the other ones can be solved by also considering the regression function $R(\cdot)$ as part of the regression problem, i.e. estimate $R(\cdot)$ from the sample such that we do not lose information about the target Y.

Dimension reduction can therefore be seen as a data preprocessing step. For example we will assume that the sample $(Y_i, \mathbf{X}_i)_{i=1}^n$ is from a model for which a linear reduction function $R(\mathbf{X}) = \mathbf{B}^T \mathbf{X}$ exists. Then the following steps can be used to tackle the regression in (1.7).

(a) Estimate the dimension k and the corresponding reduction $\mathbf{B} \in \mathbb{R}^{p \times k}$ from the data $(Y_i, \mathbf{X}_i)_{i=1}^n$ and set $\widetilde{\mathbf{X}}_i = \widehat{\mathbf{B}}^T \mathbf{X}_i \in \mathbb{R}^{\hat{k}}$

(b) Solve the original regression problem by using any regression technique for the reduced data $(Y_i, \widetilde{\mathbf{X}}_i)_{i=1}^n$

1.3 Sufficient Dimension Reduction

In this section the notion of a sufficient reduction $R(\cdot) : \mathbb{R}^p \to \mathbb{R}^k$ with $k << p$ is formalized. In the previous section we motivated the use of dimension reduction for regressions with high dimensional covariates (i.e. p being large). A sufficient reduction reduces the dimensionality of the predictor $\mathbf{X}$ at no information loss on the response Y. This will be formalized by the sufficiency of the reduction.

Definition 2. *A measurable function $R : \mathbb{R}^p \to \mathbb{R}^k$ is called a reduction function if $k \leq p$.*

Therefore any function such that the image space has lower dimension than the input space is considered as a reduction function. If $k = p$ then no real reduction takes place, so R is just a reparametrization. The classic context for dimension reduction is if p is large (i.e. $p \geq 10$) and k is relatively small (e.g., $k \leq 3$). Next we define a *sufficient* reduction for a regression.

Definition 3. *A measurable function $R : \mathbb{R}^p \to \mathbb{R}^k$ is called a sufficient reduction if and only if it is a reduction function and the response Y is conditionally independent of $\mathbf{X}$ given $R(\mathbf{X})$, i.e.*

$$Y \perp\!\!\!\perp \mathbf{X} \mid R(\mathbf{X}) \tag{1.8}$$

This definition states that a reduction function R is sufficient for Y when $R(\mathbf{X})$ contains the same information about Y as $\mathbf{X}$ does. In consequence, $R(\mathbf{X})$ is sufficient for modeling the response Y (see [Coo07]). Next a slightly different definition is presented to facilitate intuition, which we call sufficient dimension reduction model.

Definition 4. *Let the response Y be given by*

$$Y = g_{cs}(R(\mathbf{X}), \epsilon), \tag{1.9}$$

where $\mathbf{X}$ is independent of ϵ, $\epsilon \in \mathbb{R}$ is a random variable, $R(\cdot) : \mathbb{R}^p \to \mathbb{R}^k$ with $k \leq p$ is a reduction function, and g_{cs} is an unknown non-constant function.

In the model given by (1.9) the function R is a sufficient reduction since $\mathbf{X} \perp\!\!\!\perp \epsilon$, i.e. given $R(\mathbf{X})$ the response Y depends only on ϵ which is independent of $\mathbf{X}$ and definition (1.8) holds.

Theorem 1. *Suppose $(Y, \mathbf{X}^T)^T \in \mathbb{R}^{p+1}$ has a joint continuous distribution, then (1.8) and (1.9) are equivalent.*

Remark. *The proof of Theorem (1) is based on Theorem 1 of [ZZ10] who showed this Theorem for linear reductions, $R(\mathbf{X}) = \mathbf{B}^T\mathbf{X}$.*

Proof of Theorem 1. If we assume model (1.9) it is easy to see that (1.8) holds since $\epsilon \perp\!\!\!\perp \mathbf{X}$, therefore given $R(\mathbf{X})$ the response Y depends only on ϵ.

If we assume (1.8), then the definition of conditional independence (see for example [Li18] chapter 2 corollary 2.1 or [HJ17] page 460) yields

$$\begin{aligned} F_{Y|\mathbf{X}}(y|\mathbf{x}) &= \mathbb{P}\left(Y \leq y \mid \mathbf{X} = \mathbf{x}\right) = \mathbb{P}\left(Y \leq y \mid \mathbf{X} = \mathbf{x}, R(\mathbf{X}) = R(\mathbf{x})\right) \\ &= \mathbb{P}\left(Y \leq y | R(\mathbf{X}) = R(\mathbf{x})\right) = F_{Y|R(\mathbf{X})}(y|R(\mathbf{x})) \end{aligned}$$

for a fixed but arbitrary $\mathbf{x} \in \mathbb{R}^p$. Then let $\epsilon \sim U(0,1)$ be independent of $\mathbf{X}$ and set

$$\tilde{Y} = F^{-1}_{Y|R(\mathbf{X})}(\epsilon|R(\mathbf{x}))$$

where $F^{-1}_{Y|R(\mathbf{X})}(z|R(\mathbf{x}))$ is the generalized inverse (see [KR15]) of the cdf $F_{Y|R(\mathbf{X})}(y|R(\mathbf{x}))$, i.e.

$$F^{-1}_{Y|R(\mathbf{X})}(z|R(\mathbf{x})) = \inf\{y : F_{Y|R(\mathbf{X})}(y|R(\mathbf{x})) \geq z\}$$

Then we have that $(Y \mid \mathbf{X} = \mathbf{x}) = (\tilde{Y} \mid \mathbf{X} = \mathbf{x})$ in distribution by the inversion method for generating random variables, see [Kol08]. Further $\tilde{Y} = g_{\text{cs}}(R(X), \epsilon)$ with $\epsilon \perp\!\!\!\perp \mathbf{X}$ and $g_{\text{cs}}(R(X), \epsilon) = F^{-1}_{Y|R(\mathbf{X})}(\epsilon | R(\mathbf{x}))$, which completes the proof. □

An example of (1.9) is given by

$$Y = \log(1 + \mathbf{X}^T \mathbf{M} \mathbf{X})\epsilon \tag{1.10}$$

with $\mathbf{M} \in \mathbb{R}^{p \times p}$ a positive definite matrix and $\mathbf{X} \perp\!\!\!\perp \epsilon \sim N(0, 1)$. The sufficient reduction is given by the quadratic form $R(\mathbf{X}) = \mathbf{X}^T \mathbf{M} \mathbf{X} \in \mathbb{R}$ whereas a linear sufficient reduction is only possible with $k = p$, i.e. $R(\mathbf{X}) = \mathbf{I}_p \mathbf{X} = \mathbf{X}$.

Nevertheless the problem of estimating a non-linear sufficient reduction is in general quite hard or infeasible without further modeling assumptions, see [Coo07, Sec. 8.3]. For more information see [BF15] where this is considered under the assumption that the family of distributions of $\mathbf{X} \mid Y$ is known.

Using the so called kernel trick, i.e. embedding the covariate vector $\mathbf{X}$ into a higher dimensional space through a basis or kernel functions (for further information see [BJM06], [HSS08], [MRT12], and [SS18]) a non-linear sufficient reduction can be rewritten as a linear reduction. For example in the model given by (1.10) we can define a new covariate vector by adding all products $X_j X_u$ to the original covariate vector

$$\mathbf{Z} = (X_1, \ldots, X_p, X_1 X_2, X_1 X_3, \ldots, X_p X_{p-1}, X_1^2, \ldots, X_p^2)^T \in \mathbb{R}^{p + p(p+1)/2}$$

For simplicity assume that $\mathbf{M} = \mathbf{I}_p$ in model (1.10), then the regression Y onto $\mathbf{Z}$ permits a linear sufficient reduction given by $R(\mathbf{Z}) = \mathbf{B}^T \mathbf{Z} = \sum_{j=1}^{p} X_j^2 = \mathbf{X}^T \mathbf{M} \mathbf{X}$ with $\mathbf{B} = (0, 0, \ldots, 0, \boldsymbol{\iota}_p) \in \mathbb{R}^{p + p(p+1)/2}$, i.e. the vector with 0 except the last p entries containing 1. For more information about non-linear sufficient dimension reductions see [LLC13].

1.4 Linear Sufficient Dimension Reduction- the central subspace

In this section we focus on linear reductions, i.e. $R(\mathbf{X}) = \mathbf{B}^T \mathbf{X}$, that are a special case of the reductions in section 1.3 and definition (1.8). The goal of linear sufficient dimension reduction (SDR) is to find the *central subspace* $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathbb{R}^p$, i.e. a linear subspace such that the projection of the covariate vector $\mathbf{X}$ onto $\mathcal{S}_{Y|\mathbf{X}}$, i.e. $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}} \mathbf{X}$, induces no loss of information about the response Y. In [Coo98c] page 103 equation (6.3) a linear dimension reduction subspace is defined as:

Definition 5. *A linear subspace $\mathcal{S} \subseteq \mathbb{R}^p$ is called a linear dimension reduction space if for any basis $\mathbf{B} \in \mathbb{R}^{p \times k}$ of $\mathcal{S}$ the response Y is conditionally independent of $\mathbf{X}$ given $\mathbf{B}^T \mathbf{X}$, i.e.*

$$Y \perp\!\!\!\perp \mathbf{X} \mid \mathbf{B}^T \mathbf{X} \tag{1.11}$$

Then the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is defined as the intersection of all linear dimension reduction subspaces, see [Coo98c].

Definition 6. *The central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is the intersection of all linear dimension reduction subspaces,*

$$\mathcal{S}_{Y|\mathbf{X}} = \cap\{\mathcal{S} : \mathcal{S} \quad \textit{is a dimension reduction subspace}\} \tag{1.12}$$

The next question is under which assumptions a *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ exists and if it is unique. The second question is answered in Proposition 6.2 of [Coo98c], which states that if a *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ for the regression Y onto $\mathbf{X}$ exists, then it is unique. The existence is related to the question if the intersection of dimension reduction subspaces is guaranteed to be a dimension reduction subspace again. Cook answered the question negatively, i.e. without further assumptions the existence is not guaranteed, by an explicit example. Let $\mathbf{X} = (X_1, X_2)^T \in \mathbb{R}^2$ be uniformly distributed on the unit circle $\{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\| = 1\}$ and set $Y = X_1^2 + \epsilon$ for a demeaned random variable $\epsilon \perp\!\!\!\perp \mathbf{X}$. Since the support of the distribution of $\mathbf{X}$ is concentrated on the unit circle, it holds $X_1^2 + X_2^2 = 1$. Therefore $Y = X_1^2 + \epsilon = 1 - X_2^2 + \epsilon$ and we conclude that $\text{span}\{(1,0)^T\}$ and $\text{span}\{(0,1)^T\}$ are both dimension reduction subspaces but the intersection of both is $\{\mathbf{0}\}$ which is clearly not an dimension reduction subspace. In this example there exists no *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ because the two dimensional covariate vector $\mathbf{X}$ is concentrated on a one dimensional manifold. We can see this by letting $U \sim U([0, 2\pi])$ and $\mathbf{X} = (\cos(U), \sin(U))^T$, then $\mathbf{X}$ is inherently one dimensional. This means the distribution of $\mathbf{X}$ is not absolutely continuous with respect to the two dimensional Lebesgue measure since the whole probability mass of $\mathbf{X}$ is concentrated on the unit circle which has measure 0 with respect to the Lebesgue measure on $\mathbb{R}^2$.

In general, in order to guarantee the existence of $\mathcal{S}_{Y|\mathbf{X}}$, further assumptions are required. There are several ways to guarantee the existence of the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. One way would be to place assumptions on the distribution of $\mathbf{X}$. In light of the previous example, Proposition 6.4 in [Coo98c] states that if $\mathbf{X}$ has a density $f_{\mathbf{X}}$ (i.e. the distribution is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$) with convex support $\text{supp}(f_{\mathbf{X}})$, then there exists a unique $\mathcal{S}_{Y|\mathbf{X}}$.

For further discussions about the existence of $\mathcal{S}_{Y|\mathbf{X}}$ see [Coo98c, CL95, CC02].

Next we define the central subspace model

Definition 7. *Let the response Y be given by*

$$Y = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon), \tag{1.13}$$

where $\mathbf{X}$ is independent of ϵ, has a density $f_{\mathbf{X}}$ with convex support, and with a positive definite variance-covariance matrix, $\mathbb{V}\text{ar}(\mathbf{X}) = \mathbf{\Sigma}_{\mathbf{x}}$, $\epsilon \in \mathbb{R}$ is a mean zero random variable with finite $\mathbb{V}\text{ar}(\epsilon) = \mathbb{E}\left(\epsilon^2\right) = \eta^2$, g_{cs} is an unknown continuous non-constant function, and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathbb{R}^{p\times k}$ of rank $k \leq p$.

The independence of ϵ yields, that (1.13) is equivalent to

$$F(Y \mid \mathbf{X}) = F(Y \mid \mathbf{B}^T\mathbf{X}), \tag{1.14}$$

That is, Y is statistically independent of $\mathbf{X}$ when $\mathbf{B}^T\mathbf{X}$ is given and replacing $\mathbf{X}$ by $\mathbf{B}^T\mathbf{X}$ induces no loss of information for the regression of Y on $\mathbf{X}$. Furthermore in model (1.13)

it holds that

$$\operatorname{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}} \tag{1.15}$$

since, given $\mathbf{B}^T\mathbf{X}$, Y depends only on ϵ which is independent of $\mathbf{X}$. Therefore, definition (1.11) holds for $\operatorname{span}\{\mathbf{B}\}$ and, since $g_{cs} : \mathbb{R}^k \to \mathbb{R}$ is non-constant in each argument, $\operatorname{span}\{\mathbf{B}\}$ is also the smallest dimension reduction subspace. That is, for all dimension reduction sub-spaces $\mathcal{S}$ of model (1.13), $\operatorname{span}\{\mathbf{B}\} \subseteq \mathcal{S}$ and from (1.12) we can conclude (1.15). In this thesis we assume model (1.13) which implies that the *central subspace* $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$ exists and is therefore unique (see [Coo98c, Prop. 6.2 and 6.4]).

The matrix $\mathbf{B}$ in (1.13) is not identifiable and only its span is. This leads to the Grassmann manifold as the appropriate search space for $\mathcal{S}_{Y|\mathbf{X}}$. Let $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{R}$ be the QR decomposition of any matrix $\tilde{\mathbf{B}} \in \mathbb{R}^{p\times k}$ with $\operatorname{rank}(\tilde{\mathbf{B}}) = k$, i.e. $\mathbf{B} \in \mathcal{S}(p,k)$ and $\mathbf{R} \in \mathbb{R}^{k\times k}$ an upper triangular matrix. Then for any model (1.13) given by $\tilde{g}_{\text{cs}}$ and $\tilde{\mathbf{B}}$, we can write

$$Y = \tilde{g}_{\text{cs}}(\tilde{\mathbf{B}}^T\mathbf{X}, \epsilon) = \tilde{g}_{\text{cs}}(\mathbf{R}^T\mathbf{B}^T\mathbf{X}, \epsilon) = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon)$$

with $g_{cs}(\cdot,\cdot) = \tilde{g}_{\text{cs}}(\mathbf{R}^T\cdot,\cdot)$. Therefore for a given model (1.13) (i.e. $\tilde{g}_{\text{cs}}$ and $\tilde{\mathbf{B}}$) we can find an equivalent model with g_{cs} and $\mathbf{B}$ where $\mathbf{B} \in \mathcal{S}(p,k)$.

An explicit example is $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$, $\tilde{\mathbf{B}} = (c_1\mathbf{e}_1, c_2\mathbf{e}_2) \in \mathbb{R}^{p\times 2}$ with constants $c_1, c_2 \in \mathbb{R}/\{0\}$ and $\tilde{g}_{\text{cs}} : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ with $\tilde{g}_{\text{cs}}((z_1,z_2)^T, u) = z_1z_2 + u$, then we have

$$\begin{aligned} Y &= \tilde{g}_{\text{cs}}(\tilde{\mathbf{B}}^T\mathbf{X}, \epsilon) = c_1c_2X_1X_2 + \epsilon \\ &= \frac{c_1c_2}{4}\left((X_1+X_2)^2 - (X_1-X_2)^2\right) + \epsilon = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon) \end{aligned} \tag{1.16}$$

where the second equality in (1.16) is due to the polarization identity and the third by definition of $g_{cs}((z_1,z_2)^T, u) = (c_1c_2)/2\left(z_1^2 - z_2^2\right) + u$ and $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)/\sqrt{2} \in \mathcal{S}(p,2)$ with $\mathbf{b}_1 = (1,1,0,\ldots)^T \in \mathbb{R}^p$, $\mathbf{b}_2 = (1,-1,0,\ldots)^T \in \mathbb{R}^p$. Then Y can be represented by $\tilde{g}_{\text{cs}}, \tilde{\mathbf{B}}$ or equivalently by $g_{cs}, \mathbf{B}$ but $\tilde{\mathbf{B}} \neq \mathbf{B}$. Nevertheless $\operatorname{span}\{\tilde{\mathbf{B}}\} = \operatorname{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$ is unique, i.e. since in model (1.13) the function g_{cs} is unspecified and unknown we can always absorb the orthogonalization of $\mathbf{B}$ into the function g_{cs} without leaving the model class (1.13).

An interesting observation in (1.16) is that the functional form of g_{cs} depends on the particular basis of $\operatorname{span}\{\mathbf{B}\}$. This might be an interesting line of further research, since sufficient dimension reduction (SDR) is used as a first step to get an estimate $\widehat{\mathcal{S}_{Y|\mathbf{X}}}$ and afterwards the forward regression is estimated from the reduced data set $(Y_i, \widehat{\mathbf{B}}^T\mathbf{X}_i)_{i=1}^n$ where the particular basis $\widehat{\mathbf{B}}$ of $\widehat{\mathcal{S}_{Y|\mathbf{X}}}$ can be chosen arbitrarily. If one uses generalized additive models (i.e. $Y = \sum_{j=1}^k g_j(X_j) + \epsilon$ for more information see [HT90], [Woo08], and [HTFF04]) and the back-fitting algorithm to estimate the forward regression then the basis $\mathbf{B}$ and g_{cs} in (1.16) might be beneficial compared to $\tilde{\mathbf{B}}$ and $\tilde{g}_{\text{cs}}$ since the former has an additive structure whereas the latter has only an interaction term that can not be estimated by generalized additive models.

Without loss of generality, we can always assume $\mathbf{B} \in \mathcal{S}(p,k)$ in definition (1.11) and model (1.13). As only $\text{span}\{\mathbf{B}\}$ is unique, the search space for estimating $\mathcal{S}_{Y|\mathbf{X}}$ is the Grassmann manifold $Gr(p,k)$ that results in a unique parameter of interest in model (1.13). Nevertheless we will focus mostly on $\mathcal{S}(p,k)$ in this work since often it is more convenient and allows better interpretation.

With analogue computations as above, i.e. absorbing any transformation into the unknown link function, the following scaling property can be deduced (see [Coo98c, Li18]).

Theorem 2. *Assume model (1.13) holds, and let* $\mathbf{s} \in \mathbb{R}^p$, $\mathbf{O} \in \mathbb{R}^{p\times p}$ *be a non singular matrix, and set* $\mathbf{Z} = \mathbf{O}\mathbf{X} + \mathbf{s}$, *then*

$$\mathcal{S}_{Y|\mathbf{Z}} = \mathbf{O}^{-\mathbf{T}}\mathcal{S}_{Y|\mathbf{X}} \tag{1.17}$$

1.5 The mean subspace

In some cases, the problem is to estimate the *mean subspace*, denoted by $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. Imposing an additiver error term in model (1.13), yields the mean subspace model.

Definition 8. *Suppose* $(Y, \mathbf{X}^T)^T$ *has a joint continuous distribution, where* $Y \in \mathbb{R}$ *denotes a univariate response and* $\mathbf{X} \in \mathbb{R}^p$ *a p-dimensional covariate vector. We assume that the dependence of* Y *and* $\mathbf{X}$ *is modelled by*

$$Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon, \tag{1.18}$$

where $\mathbf{X}$ *is independent of* ϵ, *has a density* $f_{\mathbf{X}}$ *with convex support, with positive definite variance-covariance matrix,* $\mathbb{V}\text{ar}(X) = \mathbf{\Sigma}_{\mathbf{x}}$, $\epsilon \in \mathbb{R}$ *is a mean zero random variable with finite* $\mathbb{V}\text{ar}(\epsilon) = E\left(\epsilon^2\right) = \eta^2 < \infty$, g *is an unknown continuous non-constant function, and* $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathbb{R}^{p\times k}$ *of rank* $k \leq p$.

In model (1.18), $\text{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$, so that the central subspace agrees with the mean subspace, because of the independence of $\mathbf{X}$ and ϵ. If the independence requirement is replaced with $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$, we still have $\text{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ due to

$$\mathbb{E}\left(Y \mid \mathbf{X}\right) = \mathbb{E}\left(g(\mathbf{B}^T\mathbf{X}) \mid \mathbf{X}\right) + \underbrace{\mathbb{E}\left(\epsilon \mid \mathbf{X}\right)}_{=0} = g(\mathbf{B}^T\mathbf{X}) = \mathbb{E}\left(Y \mid \mathbf{B}^T\mathbf{X}\right). \tag{1.19}$$

but the mean subspace can be a proper subset of the central subspace, see example (1.22) next. A distribution is not uniquely determined by its first moment, therefore it always holds

$$\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} \tag{1.20}$$

The *mean subspace* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ captures all the information in $\mathbf{X}$ about Y that is contained in the first conditional moment $\mathbb{E}(Y \mid \mathbf{X})$, i.e. if we are only interested in the conditional mean, it suffices to know $\mathbf{B}^T\mathbf{X} \in \mathbb{R}^k$ instead of $\mathbf{X} \in \mathbb{R}^p$.

Another characterisation analogue to (1.11) is given by: For any basis $\mathbf{B} \in \mathbb{R}^{p\times k}$ of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$

$$\mathbb{E}\left(Y \mid \mathbf{X}\right) = E(Y \mid \mathbf{B}^T\mathbf{X}) \tag{1.21}$$

Example The following model highlights the difference between the *mean subspace* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ and *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. Let, $p = 10$, $k = 2$ $\mathbf{X} = (X_1, \ldots, X_{10})^T \sim N(\mathbf{0}, \mathbf{I}_{10})$, $\epsilon \sim N(0,1)$ independent to $\mathbf{X}$, $\mathbf{b}_1 = \mathbf{e}_1 = (1,0,0,\ldots)^T \in \mathbb{R}^{10}$, $\mathbf{b}_2 = \mathbf{e}_2 = (0,1,0,0,\ldots)^T \in \mathbb{R}^{10}$, $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2) \in \mathbb{R}^{10\times 2}$, and the response Y is given by

$$Y = (\mathbf{b}_1^T\mathbf{X})^2 + (|\mathbf{b}_2^T\mathbf{X}| + 1)\epsilon = X_1^2 + (|X_2| + 1)\epsilon \tag{1.22}$$

In this model the *mean subspace* is a proper subset of the *central subspace*, i.e. $\operatorname{span}\{\mathbf{b}_1\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subset \mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$. To see this we calculate

$$\mathbb{E}\left(Y \mid \mathbf{X}\right) = \mathbb{E}\left((\mathbf{b}_1^T\mathbf{X})^2 \mid \mathbf{X}\right) + \underbrace{\mathbb{E}\left((|\mathbf{b}_2^T\mathbf{X}| + 1)\epsilon \mid \mathbf{X}\right)}_{=(|\mathbf{b}_2^T\mathbf{X}|+1)\mathbb{E}(\epsilon|\mathbf{X})=0} = (\mathbf{b}_1^T\mathbf{X})^2 = \mathbb{E}\left(Y \mid \mathbf{b}_1^T\mathbf{X}\right)$$

where the measurablility of $(|\mathbf{b}_2^T\mathbf{X}| + 1)$ is with respect to the sigma algebra generated by $\mathbf{X}$, and the independence of ϵ and $\mathbf{X}$ was used for the term that vanished. Further we conclude $Y - \mathbb{E}\left(Y \mid \mathbf{X}\right) = (|\mathbf{b}_2^T\mathbf{X}| + 1)\epsilon$ and the conditional variance is given by

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(Y \mid \mathbf{X}\right) &= \mathbb{E}\left((Y - \mathbb{E}\left(Y \mid \mathbf{X}\right))^2 \mid \mathbf{X}\right) = \mathbb{E}\left((|\mathbf{b}_2^T\mathbf{X}| + 1)^2\epsilon^2 \mid \mathbf{X}\right) \\ &= (|\mathbf{b}_2^T\mathbf{X}| + 1)^2 \underbrace{\mathbb{E}(\epsilon^2 \mid \mathbf{X})}_{=\mathbb{E}(\epsilon^2)=\eta^2} = (|\mathbf{b}_2^T\mathbf{X}| + 1)^2\eta^2\end{aligned}$$

where the measurability of $(|\mathbf{b}_2^T\mathbf{X}| + 1)^2$ with respect to the sigma algebra generated by $\mathbf{X}$, and the independence of ϵ and $\mathbf{X}$ were used. Moreover the response Y depends only on $\mathbf{B}^T\mathbf{X} = (X_1, X_2)^T \in \mathbb{R}^2$ so the *central subspace* is given by $\mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$.

To summarize, model (1.22) is a regression problem with heteroskedastic errors, i.e. the conditional first moment $\mathbb{E}\left(Y \mid \mathbf{X}\right) = X_1^2$ is determined by $\mathbf{b}_1^T\mathbf{X}$ and the conditional variance $\mathbb{V}\mathrm{ar}\left(Y \mid \mathbf{X}\right) = (|X_2| + 1)^2\eta^2$is determined by $\mathbf{b}_2^T\mathbf{X}$.

1.6 A historic account of Sufficient Dimension Reduction

Albeit dimension reduction (i.e. principal component analysis and similar concepts) was used in a variety of fields before, systematic studies of sufficient dimension reduction (SDR) in a probabilistic context are, from a historic point of view, quiet novel since the theoretical foundation of this sub-field of statistics was developed around 1991 by the introduction of SIR [Li91]. To quote R. Dennis Cook from the paper [Coo18]: *Interpreted broadly, dimension reduction has always been a bedrock of statistical thought.*

To be precise, ordinary least square estimation, the power house of statistics, can also be viewed as sufficient dimension reduction (SDR) technique. To see this, let $g_{\text{cs}} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$

with $g_{cs}(z,u) = \alpha + bz + u$ with $\alpha, b \in \mathbb{R}$ and $\mathbf{B} \in \mathbb{R}^p$, $\|\mathbf{B}\| = 1$ (i.e. the dimension of the central subspace is $\dim(\mathcal{S}_{Y|\mathbf{X}}) = k = 1$), then model (1.13) can be written as

$$Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon = \alpha + \underbrace{b\mathbf{B}^T}_{=\boldsymbol{\beta}}\mathbf{X} + \epsilon = \alpha + \boldsymbol{\beta}^T\mathbf{X} + \epsilon \tag{1.23}$$

the standard model that is assumed for ordinary least square estimation. Therefore we note that sufficient dimension reduction is a generalization to the linear regression set-up. In model (1.23) ordinary least square (OLS) estimation of $\boldsymbol{\beta}$ performs estimation of the linear link function and of the sufficient dimension reduction subspace $\text{span}\{\mathbf{B}\} = \text{span}\{\boldsymbol{\beta}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$ simultaneously. Therefore the ordinary least square estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}}^T)^T \in \mathbb{R}^{p+1}$ estimates the parameters α, b of the linear link function through $\hat{\alpha}, \|\hat{\boldsymbol{\beta}}\|$ and the SDR subspace $\text{span}\{\mathbf{B}\}$ by $\text{span}\{\hat{\boldsymbol{\beta}}\}$. Nevertheless under the assumption of a linear link function we restrict the dimension of central subspace $\mathcal{S}_{Y|\mathbf{X}}$ (i.e. which equals the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ in model (1.23), see section 1.5) to be 1, i.e. $\dim(\mathcal{S}_{Y|\mathbf{X}}) = \text{rank}(\mathbf{B}) = k = 1$. Therefore model (1.23) is to restrictive and narrow to study sufficient dimension reduction.

The next step in the direction of sufficient dimension reduction came from Brillinger (1977, 1983) where he noted that ordinary least square estimation (OLS, multiple regression, etc ....), where the link function g in (1.18) that links the response Y to the predictors $\mathbf{X}$ is assumed to be linear, had been applied successfully in a very wide variety of scientific fields. In his view it was too successful for the narrow assumptions placed on the probabilistic model to derive the ordinary least squares (OLS) estimator. Especially he noted that often the researchers using ordinary least square estimation were not interested in the actual values of the estimated coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T \in \mathbb{R}^p$ but instead drew conclusion from the relative coefficients, i.e. $\hat{\beta}_j/\hat{\beta}_k$ for $1 \leq j, k \leq p$. This led him to investigate the OLS estimator under link violation (i.e. how does ordinary least square estimation perform under the presence of a nonlinear link function) in the paper [Bri12], i.e. he studied the model given by

$$Y = g(\alpha + b\mathbf{B}^T\mathbf{X}) + \epsilon = g(\alpha + \boldsymbol{\beta}^T\mathbf{X}) + \epsilon \tag{1.24}$$

with $\boldsymbol{\beta} = b\mathbf{B} \in \mathbb{R}^p$. Model (1.24) is also called single index model (for further information see [Ich93, HHI93, Rad15, HS89]) and is again a special case of model (1.13) with $g_{cs} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ and $g_{cs}(z,u) = g(\alpha + bz) + u$ where $\alpha, b \in \mathbb{R}$ and $\mathbf{B} \in \mathbb{R}^p$, $\|\mathbf{B}\| = 1$. In (1.24) the dimension of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is again $1 = \dim(\text{span}\{\mathbf{B}\}) = \dim(\text{span}\{\boldsymbol{\beta}\}) = \dim(\mathcal{S}_{Y|\mathbf{X}})$ as was the case in the ordinary least squares model in (1.23). His question was: How does ordinary least square estimation perform under the presence of a nonlinear link function g in (1.24). In Theorem 1 in [Bri12] it is shown that under the assumption that $\mathbf{X} \perp\!\!\!\perp \epsilon$ and $\mathbf{X} \sim N(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$ the ordinary least square estimator $\hat{\boldsymbol{\beta}}$ from an independent identical distributed sample of model (1.24) converges in probability to a multiple of the population quantity, i.e.

$$\hat{\boldsymbol{\beta}} \to c\boldsymbol{\beta} \quad \text{in probability as the sample size } n \text{ goes to infinity} \tag{1.25}$$

for some $c \in \mathbb{R}$. This result answered the question why the ordinary least squares estimator is so successful even if the link function is miss-specified, especially when it is used to drew

conclusions from the relative coefficients $\hat{\beta}_j/\hat{\beta}_k \to \beta_j/\beta_k$ since they converge in probability to the true ratios if the constant c in (1.25) is nonzero (see model (1.28)). Further, from a sufficient dimension reduction viewpoint, it shows that asymptotically the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ consistently estimates one direction of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, i.e. $\text{span}\{\hat{\boldsymbol{\beta}}\} \subset \mathcal{S}_{Y|\mathbf{X}}$ as the sample size n goes to infinity.

Duan and Li in [LD89] extended the result of [Bri12] to the model given by

$$Y = g(\alpha + \boldsymbol{\beta}^T\mathbf{X}, \epsilon) \tag{1.26}$$

which is again a special case of model (1.13) with $g_{\text{cs}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $g_{\text{cs}}(z, u) = g(\alpha + bz, u)$ where $\alpha, b \in \mathbb{R}$ and $\boldsymbol{\beta} = b\mathbf{B} \in \mathbb{R}^p$. Model (1.26) is a generalization of (1.24) without an additive error structure where the one dimensional central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is spanned by $\boldsymbol{\beta}$, i.e. $\text{span}\{\boldsymbol{\beta}\} = \text{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$ with $\dim(\text{span}\{\boldsymbol{\beta}\}) = k = 1$. Theorem 2.1 of [LD89] states that under model (1.26) and the so called *linearity condition*

$$\mathbb{E}\left(\mathbf{b}^T\mathbf{X}|\boldsymbol{\beta}^T\mathbf{X}\right) \text{ is linear in } \boldsymbol{\beta}^T\mathbf{X} \text{ for all } \mathbf{b} \in \mathbb{R}^p \tag{1.27}$$

the statement in (1.25) holds for some $c \in \mathbb{R}$. The *linearity condition* given in (1.27) holds for the normal distribution and for distributions in the elliptically contoured family (i.e. if (1.27) holds for all $\boldsymbol{\beta}$ then $\mathbf{X}$ has an elliptically contoured distribution and the other way round, see [Eat86] and [Li18] page 14 and chapter 7 corollary 7.1). The distribution of a random vector $\mathbf{X} \in \mathbb{R}^p$ falls in the elliptically contoured family if its characteristic function $\phi_{\mathbf{X}-\boldsymbol{\mu}}(\mathbf{b}) = \mathbb{E}(\exp(i\mathbf{b}^T(\mathbf{X} - \boldsymbol{\mu})))$ fulfills the functional equation $\phi_{\mathbf{X}-\boldsymbol{\mu}}(\mathbf{b}) = \psi(\mathbf{b}^T\boldsymbol{\Sigma}\mathbf{b})$ for all $\mathbf{b} \in \mathbb{R}^p$, for some location parameter $\boldsymbol{\mu} \in \mathbb{R}^p$, some positive definite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$, and some function ψ. Another point of view on elliptically contoured distributions is that the contour lines of the density are elliptical. For more information about the elliptically contoured distributions see [CHS81].

Remark. *Note that (1.27) must be fulfilled for the $\boldsymbol{\beta}$ given by model (1.26) but since in practice we do not know $\boldsymbol{\beta}$ we will require that (1.27) holds for all possible $\boldsymbol{\beta}$. The linearity condition or linear design condition (i.e. if (1.27) is extended to $\beta \in \mathcal{S}(p, k)$) in (1.27) will be quite important later on in the history of sufficient dimension reduction techniques and should be viewed as a weakening of the assumption that the predictors $\mathbf{X}$ are normal distributed. Nevertheless it is still a restrictive assumption that rules out most of the distributions for $\mathbf{X}$ but [DF84] showed that when $\mathbf{X}$ is high dimensional, projections of $\mathbf{X}$ are approximately normal distributed. Furthermore in [HL93] it is shown that if $\mathbf{X}$ is high dimensional, then $\mathbb{E}(\mathbf{X} \mid \mathbf{B}^T\mathbf{X})$ is approximately linear in the conditioning argument if $\mathbf{B} \in \mathcal{S}(p, k)$ with $k << p$. Moreover [SL18] showed that the conditional expectation in (1.27) is linear for $p \to \infty$ (i.e. $\mathbf{X} \in \mathbb{R}^p$ bein high dimensional) if $\mathbf{X}$ has a density with respect to the Lebesgue-measure, certain moments are close to the Gaussian moments, and some further technical integrability conditions are satisfied. These statements justifies the assumption of the linear design condition (LDC) since $\boldsymbol{\beta}^T\mathbf{X}$ and $\mathbf{b}^T\mathbf{X}$ can be regarded as projections. Furthermore $p >> k$ (i.e. p being much larger than k and k being quite small) is exactly the set-up that linear sufficient dimension reduction (linear SDR) is about, since we want to replace a high dimensional covariate vector $\mathbf{X}$ by a much lower dimensional projection.*

Remark. *Nevertheless there is the caveat that the proportionality constant c of the result (1.25) in models (1.24) and (1.26) can be 0, i.e. asymptotically* $\operatorname{span}\{\hat{\boldsymbol{\beta}}\} = \{\mathbf{0}\}$. *To see this, let* $\mathbf{X} = (X_1, \dots, X_p)^T \sim N(\mathbf{0}, \mathbf{I}_p)$, $\epsilon \sim N(0,1)$ *independent of* $\mathbf{X}$, $\boldsymbol{\beta} = \mathbf{e}_1 \in \mathcal{S}(p,1)$, *and let*

$$Y = g(\boldsymbol{\beta}^T\mathbf{X}) + \sigma(\boldsymbol{\beta}^T\mathbf{X})\epsilon \tag{1.28}$$

for a even function $g : \mathbb{R} \to \mathbb{R}$ (i.e $g(z) = g(-z)$) and a measurable function $\sigma : \mathbb{R} \to \mathbb{R}$. This model fits into model (1.26) and for the constant function $\sigma(\cdot) = \sigma$ the model is a special case of (1.24). In both configurations model (1.28) fulfills all assumptions such that (1.25) holds. Let $\mathcal{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ be an i.i.d. sample of model (1.28), then the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ converges asymptotically to

$$\hat{\boldsymbol{\beta}} = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T\mathbf{Y} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n \mathbf{X}_iY_i \longrightarrow \tag{1.29}$$
$$\underbrace{\mathbb{V}\mathrm{ar}\,(\mathbf{X})^{-1}}_{=\mathbf{I}_p}\, cov\,(\mathbf{X}, Y) = cov\,(\mathbf{X}, Y) \quad \text{almost surely as } n \to \infty$$

since by the strong law of large numbers and $\mathbb{E}(\mathbf{X}) = 0$ it holds $\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n \to \mathbb{V}\mathrm{ar}(\mathbf{X})$ and $\sum_{i=1}^n \mathbf{X}_iY_i/n \to cov(\mathbf{X}, Y)$ almost surely. The covariance is given by

$$\begin{aligned} cov\,(\mathbf{X}, Y) &= cov\left(\mathbf{X}, g(\boldsymbol{\beta}^T\mathbf{X})\right) + cov\left(\mathbf{X}, \sigma(\boldsymbol{\beta}^T\mathbf{X})\epsilon\right) \\ &= \underbrace{\mathbb{E}\left(\mathbf{X}g(\boldsymbol{\beta}^T\mathbf{X})\right)}_{(\mathbb{E}(X_1g(X_1)),\mathbb{E}(X_2g(X_2)),\dots,\mathbb{E}(X_pg(X_1)))} + \underbrace{\mathbb{E}\left(\mathbf{X}\sigma(\boldsymbol{\beta}^T\mathbf{X})\epsilon\right)}_{\mathbb{E}\left(\mathbf{X}\sigma(\boldsymbol{\beta}^T\mathbf{X})\mathbb{E}(\epsilon|\mathbf{X})\right)} = \mathbf{0} \end{aligned} \tag{1.30}$$

where the first equality is by inserting model (1.28), the second by $\mathbb{E}(\mathbf{X}) = 0$. The second term in (1.30) vanishes due to $\mathbb{E}(\epsilon \mid \mathbf{X}) = E(\epsilon) = 0$ by $\mathbf{X} \perp\!\!\!\perp \epsilon$. The first term in (1.30) vanishes since the first component $X_1 \sim N(0,1)$ has an even density on a symmetric domain and $h(z) = zg(z)$ being odd, therefore $\mathbb{E}(X_1g(X_1)) = 0$ and for the other components

$$\mathbb{E}(X_jg(X_1)) = \underbrace{\mathbb{E}(X_j)}_{=0}\mathbb{E}(g(X_1)) = 0 \quad \text{for} \quad 2 \le j \le p$$

due to the independence of the components of a multivariate normal if the covariance is $\mathbf{I}_p$, which yields that the first term in (1.30) is 0.

Therefore we have $\hat{\boldsymbol{\beta}} \to \mathbf{0} \neq \boldsymbol{\beta}$ almost surely and we can conclude that the proportionality constant c in (1.25) must be 0.

The next step in the direction of modern sufficient dimension reduction (SDR) was the development of projection pursuit regression methods which is a generalization of generalized additive models (for further details see [HT90], [Woo08], and [HTFF04]). Projection pursuit regression is again a special case of the model (1.13) where $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_r) \in \mathbb{R}^{p\times r}$ with $\|\mathbf{b}_j\| = 1$ and $\mathbf{g}_{\text{cs}} : \mathbb{R}^r \times \mathbb{R} \to \mathbb{R}$ $\mathbf{g}_{\text{cs}}(\mathbf{B}^Tz, u) = \sum_{j=1}^r g_j(\mathbf{b}_j^Tz) + u$ with $z \in \mathbb{R}^p$ and

functions $g_j : \mathbb{R} \to \mathbb{R}$. The model for projection pursuit regression is given by

$$Y = \mathbf{g}_{\text{cs}}(\mathbf{B}^T\mathbf{X}, \epsilon) = \sum_{j=1}^{r} g_j(\mathbf{b}_j^T\mathbf{X}) + \epsilon \tag{1.31}$$

where the functions g_j and $\mathbf{b}_j$ are estimated recursively, for more information see [FS81], [HTFF04], and [KG12]. Nevertheless the intend of projection pursuit regression is only to estimate the link function g_{cs} in model (1.13) if an additive error structure is assumed by a superposition of r function and not to find lower dimensional projections $\mathbf{B}^T\mathbf{X}$ that induce no loss of information, i.e. the number of projections r is allowed to exceed p. That is r serves as a complexity parameter of the model and for $r \to \infty$ model (1.31) becomes a universal approximator for functions with domain $\mathbb{R}^p$ (i.e. $\lim_{r\to\infty} \sum_{j=1}^{r} g_j(\mathbf{b}_j^T z) \approx g(z)$ for any function $g : \mathbb{R}^p \to \mathbb{R}$). In practice the number r is chosen by cross-validation or a forward-step-wise criterion and often r is determined to be quite low, i.e. $r << p$, so that projection pursuit regression performs dimension reduction nevertheless.

1.7 Overview of SDR methods

The first method targeting the linear sufficient reduction in the general regression model $F(Y \mid \mathbf{X}) = F(Y \mid \mathbf{B}^T\mathbf{X})$, where F signifies the conditional cumulative distribution function of Y given the conditioning argument, was *sliced inverse regression* (SIR, [Li91]). SIR, as well as most *sufficient dimension reduction* (SDR) methods, is based on the *inverse regression* of $\mathbf{X}$ on the response Y. These include *sliced average variance estimation* (SAVE, [Coo00]), *parametric inverse regression* (PIR, [BC01]), *principal fitted components* (PFC, [CF08]), *directional regression* (DR, [LW07a]), and *contour regression* (CR, [LZC05a]). Further, there are model (likelihood) based sufficient dimension reduction methods, overlapping with the inverse regression based methods, such as *likelihood acquired directions* (LAD, [CF09]), which mostly require assumptions on the conditional $\mathbf{X} \mid Y$ or joint distribution and are researched in [Coo07, CF09, BF15, BDF16]. A recent overview of SDR methods can be found in [Yin, MZ13, Li18].

These methods require varying assumptions on either the joint distribution of $(Y, \mathbf{X}^T)^T$, or the conditional distribution of $\mathbf{X} \mid Y$, limiting their applicability. A different approach focuses on the *forward* regression of Y on $\mathbf{X}$ in order to extract the reduction. The first such method, *principal Hessian directions* (pHd), was introduced by [Li92] and was further developed by [Coo98a] and [CL02, CL04]. *Minimum average variance estimation* (MAVE) was introduced by [XTLZ02] and was generalized in [Xia07, WX08]. *Conditional variance estimation* (CVE in chapter 2, [FB21a]) is the most recent addition to the forward regression SDR methodology. These estimators require minimal assumptions on the smoothness of the joint distribution and frequently enjoy better estimation accuracy but at the expense of higher computational cost. Among those, the most prominent so far has been the *minimum average variance estimation* (MAVE) [XTLZ02].

1.8 Inverse Regression

The first method, *sliced inverse regression* (SIR), that can be characterized as a pure sufficient dimension reduction (SDR) technique based on a probabilistic model, i.e. model (1.13), with the intend of estimating $\mathcal{S}_{Y|\mathbf{X}}$ was introduced by Ker-Chau Li in 1991 in his famous paper [Li91]. *Sliced inverse regression* (SIR) is based on the so called inverse regression problem where we regress $\mathbf{X}$ on Y, i.e. $\mathbb{E}(\mathbf{X}|Y)$. Theorem 3 gives the explanation why the inverse regression is useful in the dimension reduction context.

Theorem 3. *Assume model* (1.13) *and the linear design condition (LDC) in* (1.27)*, i.e.* $\mathbb{E}(\mathbf{X} \mid \mathbf{B}^T\mathbf{X})$ *is linear in the conditioning argument, where* $\mathbf{B}$ *is given in* (1.13)*, then*

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} \operatorname{span}\{\mathbb{E}(\mathbf{X}|Y) - \mathbb{E}(\mathbf{X})\} \subseteq \mathcal{S}_{Y|\mathbf{X}} \quad \textit{almost surely} \tag{1.32}$$

Proof of Theorem 3. Let $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \mathbb{E}(\mathbf{X}))$ be the standardized predictors (i.e. $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}$ is the inverse of the symmetric matrix root of $\mathbb{V}\mathrm{ar}(\mathbf{X}) = \boldsymbol{\Sigma}_{\mathbf{x}}$) and by the tower property of the conditional expectation, $\epsilon \perp\!\!\!\perp \mathbf{X}$, and the linear design condition (LDC), it holds

$$\mathbb{E}(\mathbf{Z}|Y) = \mathbb{E}\left(\mathbb{E}(\mathbf{Z} \mid \mathbf{B}^T\mathbf{Z}, \epsilon)|Y\right) = \mathbb{E}\left(\underbrace{\mathbb{E}(\mathbf{Z} \mid \mathbf{B}^T\mathbf{Z})}_{=\mathbf{A}\mathbf{B}^T\mathbf{Z}}|Y\right) = \mathbf{A}\mathbf{B}^T\mathbb{E}(\mathbf{Z}|Y) \tag{1.33}$$

where all equalities are in almost sure sense. Then the projection Theorem for Hilbert spaces yields

$$\mathbf{A} = \operatorname{cov}(\mathbf{Z}, \mathbf{B}^T\mathbf{Z})\mathbb{V}\mathrm{ar}(\mathbf{B}^T\mathbf{Z})^{-1} = \mathbb{V}\mathrm{ar}(\mathbf{Z})\mathbf{B}\left(\mathbf{B}^T\mathbb{V}\mathrm{ar}(\mathbf{Z})\mathbf{B}\right)^{-1} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1} \tag{1.34}$$

since $\mathbb{V}\mathrm{ar}(\mathbf{Z}) = \mathbf{I}_p$ and inserting (1.34) into (1.33) yields

$$\mathbb{E}(\mathbf{Z}|Y) = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbb{E}(\mathbf{Z}|Y) = \mathbf{P}_{\mathbf{B}}\mathbb{E}(\mathbf{Z}|Y)$$

and we conclude $\mathbb{E}(\mathbf{Z}|Y) \in \mathcal{S}_{Y|\mathbf{Z}}$ and by Theorem 2 it holds $\mathcal{S}_{Y|\mathbf{Z}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{Y|\mathbf{X}}$. Therefore

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y) \in \mathcal{S}_{Y|\mathbf{X}} \quad \text{almost surely}$$

□

Corollary 4. *Under the assumptions of Theorem 3, it holds*

$$\operatorname{span}\{\mathbb{V}\mathrm{ar}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right)\} = \operatorname{span}\{\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y = y) : \quad y \in \mathit{supp}(Y)\}$$

Proof. Let $\mathbf{V} \in \mathcal{S}(p, k)$ be a orthonormal basis of a linear subspace $\mathcal{S} \subseteq \mathbb{R}^p$, i.e. $\operatorname{span}\{\mathbf{V}\} = \mathcal{S}$. If a random variable $\mathbf{Z} \in \operatorname{span}\{\mathbf{V}\} \subseteq \mathbb{R}^p$ almost surely, then it holds that

$$\mathbf{Z} = \mathbf{P}_{\operatorname{span}\{\mathbf{V}\}}\mathbf{Z} \quad \text{almost surely}$$

since $\mathbf{P}_{\text{span}\{\mathbf{V}\}^\perp}\mathbf{Z} = \mathbf{0}$ almost surely. Therefore

$$\mathbf{Z} = \mathbf{P}_{\text{span}\{\mathbf{V}\}}\mathbf{Z} = \mathbf{V}\underbrace{\mathbf{V}^T\mathbf{Z}}_{=\tilde{\mathbf{Z}}} = \mathbf{V}\tilde{\mathbf{Z}} \quad \text{almost surely}$$

and applying the variance operator on both side yields

$$\mathbb{V}\text{ar}(\mathbf{Z}) = \mathbf{V}\mathbb{V}\text{ar}(\tilde{\mathbf{Z}})\mathbf{V}^T,$$

i.e. $\text{span}\{\mathbb{V}\text{ar}(\mathbf{Z})\} = \text{span}\{\mathbf{V}\}$. Setting $\mathbf{Z} = \mathbb{E}(\mathbf{X}-\mathbb{E}(\mathbf{X})|Y)$ and $\mathcal{S} = \text{span}\{\mathbb{E}(\mathbf{X}-\mathbb{E}(\mathbf{X})|Y = y) : \quad y \in \text{supp}(Y)\}$ (which is a linear subspace by definition of the span and therefore permits a basis $\mathbf{V}$ of $\mathcal{S}$) yields the result. □

Theorem 3 and corrolary 4 are the foundation of most first order sufficient dimension reduction (SDR) methods that relay on the inverse problem (i.e. regressing $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ on Y and the name first order inverse regression refers to the fact that only the first conditional moment of the inverse regression is used). They state that the inverse regression function $\mathbb{E}(\mathbf{X}|Y)$ ranges in the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, and therefore given an estimate for $\mathbb{E}(\mathbf{X}|Y)$ and its variance, we can estimate a part of $\mathcal{S}_{Y|\mathbf{X}}$. This has the advantage that it circumvents part of the course of dimensionality since the inverse regression problem consists actually of p one dimensional regression problems, i.e. $\mathbb{E}(X_j|Y)$ for $j = 1, \ldots, p$. Nevertheless there is the caveat that there is no guarantee that one can estimate the whole central subspace $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively since Theorem 3 states only that $\mathbb{V}\text{ar}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right)$ spans a subset of $\mathcal{S}_{Y|\mathbf{X}}$. In the extreme case, i.e. if $\mathbb{E}(\mathbf{X}|Y)$ is constant with respect to Y, we would have $\mathbb{V}\text{ar}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right) = \mathbf{0} \in \mathbb{R}^{p\times p}$. To see this, consider the example in the next remark.

Remark. *Let $\mathbf{U} = (U_1, \ldots, U_p)^T \sim U([-1, 1]^p)$ be uniformly distributed on the p dimensional cube with independent components, and set $Y = U_1^2$, i.e. $\text{span}\{(1, 0, \ldots)^T\} = \mathcal{S}_{Y|\mathbf{X}}$. Then due to independence it holds $\mathbb{E}(U_j|Y) = \mathbb{E}(U_j) = 0$ for $j = 2, \ldots, p$ and calculate*

$$\mathbb{E}(U_1|Y = y) = \mathbb{E}\left(-\sqrt{y}1_{\{U_1<0\}} + \sqrt{y}1_{\{U_1>0\}}\right) = (-\sqrt{y} + \sqrt{y})0.5 = 0.$$

Therefore $\mathbb{E}(\mathbf{X}|Y) = \mathbf{0}$ and $\mathbb{V}\text{ar}(\mathbb{E}(\mathbf{X}|Y)) = \mathbf{0} \in \mathbb{R}^{p\times p}$. This example can be generalized to even link functions and symmetric distributions, therefore inverse regression methods based on Theorem 3 cannot estimate the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively in such settings.

1.9 Sliced Inverse Regression (sir)

Assume $(Y_i, \mathbf{X}_i)_{i=1}^n$ is an i.i.d. sample from model (1.13) and assume the linear design condition (LDC) given by (1.27). Then the SIR [Li91] algorithm is given by

(a) Standardize $\mathbf{Z}_i = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$ where $\bar{\mathbf{X}} = (1/n)\sum_{i=1}^n \mathbf{X}_i$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} = (1/n)\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$.

(b) Divide the range of Y into $H \in \mathbb{N}$ slices $I_h = (q_{h-1}, q_h]$ for $h = 1, \dots, H$ (often empirical quantiles are used for q_h such that there are approximately equal number of Y_i in each slice) and let $n_h = \#\{i : Y_i \in I_h\}$, i.e. the number of observations Y_i that fall into slice I_h

(c) Calculate the averages $\bar{\mathbf{Z}}_h = (1/n_h) \sum_{i:Y_i \in I_h} \mathbf{Z}_i \approx \mathbb{E}(\mathbf{Z}|Y \in I_h)$

(d) Calculate the empirical variance matrix of $\bar{\mathbf{X}}_h$, i.e.

$$\widehat{\mathbf{M}}_1 = \sum_{h=1}^{H} (n_h/n) \bar{\mathbf{Z}}_h \bar{\mathbf{Z}}_h^T \approx \mathbb{V}\mathrm{ar}\left(\mathbb{E}(\mathbf{Z}|Y)\right)$$

(e) Calculate the spectral decomposition of $\widehat{\mathbf{M}}_1$ and set $\widehat{\mathbf{B}} = \widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1/2}(\hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_k)$ where $\hat{\mathbf{b}}_j$ for $j = 1, \dots, k$ are the eigenvectors of $\widehat{\mathbf{M}}_1$ corresponding to the k largest eigenvalues

Remark. *The number of slices H and the intervals I_h are tuning parameters but [Li91] showed that the algorithm estimates* $\mathrm{span}\{\mathbb{V}\mathrm{ar}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right)\}$ *consistently with rate $1/\sqrt{n}$ no matter how H and I_h is chosen (even in the worst case if there are only two points in each interval) but the asymptotic variance may be inflated by a bad choice of the tuning parameters. A reasonable choice proposed by Li is $I_h = (F_Y^{-1}((h-1)/H), F_Y^{-1}(h/H)]$ where F_Y denotes the cumulative distribution function (cdf) of Y and for the implementation we use the empirical distribution function $F_{Y,n}(y) = (1/n) \sum_{i=1}^{n} 1_{\{Y_i \le y\}}$ instead of F_Y. This corresponds to choosing q_h as the empirical quantiles.*

Further, note that the dimension $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$ of the central subspace is assumed to be known. In practice this is not the case and there are test for the dimension k based on asymptotic results. The most widely used one, due to its simplicity, is based on the fact that if $\mathbf{X}$ is normal distributed then $n(p-k)\bar{\lambda}_{p-k}$ follows a chi-squared distribution χ^2 with $(p-k)(H-k-1)$ degrees of freedom (see [Li91]) where $\bar{\lambda}_{p-k} = (1/(p-k)) \sum_{j=k+1}^{p} \lambda_j$ with λ_j the eigenvalues in descending order of $\widehat{\mathbf{M}}_1$ in the SIR *algorithm; i.e. under the null hypothesis that model (1.13) with $k = rank(\mathbf{B})$ holds the average of the $p - k$ smallest eigenvalues (their population counterparts are 0 due to Theorem 3 and corrolary 4) is asymptotically chi-squared distributed. This can be used to perform an sequential test procedure for the dimension $k = rank(\mathbf{B}) = \dim(\mathcal{S}_{Y|\mathbf{X}})$ in model (1.13). For further methods of estimating k see [Coo98c], [BC01].*

Remark. *Note that if $H = n$, i.e. the number of slices H equals the sample size n such that in each slice I_h there is only one point Y_i, then Sliced Inverse Regression recovers the principal components.*

Next a reformulation of the SIR algorithm in terms of a generalized eigenvalue (1.4) problem is presented.

Remark. *Note that Theorem 3 and corrolary 4 can be formulated as a generalized eigenvalue problem. Let* $\mathbf{M}_1 = \mathbb{V}\mathrm{ar}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right)$*, then* $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\,\mathrm{span}\{\mathbf{M}_1\} = \mathrm{span}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{M}_1\} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \mathrm{span}\{\mathbf{B}\}$ *with* $\dim(\mathcal{S}_{Y|\mathbf{X}}) = rank(\mathbf{B}) = k$*, therefore* $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{M}_1$ *has at most* k *eigenvectors* $\mathbf{v}_j$ *with eigenvalues* $\lambda_j \neq 0$ *(for simplicity assume for the moment that there exist exactly* k *eigenvectors with eigenvalue not equal to* 0*, i.e. assume exhaustiveness* $\mathrm{span}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{M}_1\} = \mathcal{S}_{Y|\mathbf{X}}$*). Then* $\mathrm{span}\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{M}_1\} = \mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \mathcal{S}_{Y|\mathbf{X}}$ *and*

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{M}_1\mathbf{v} = \lambda\mathbf{v} \quad \Longleftrightarrow \quad \mathbf{M}_1\mathbf{v} = \lambda\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{v} \quad \textit{for} \quad j \leq k \tag{1.35}$$

Moreover the final step (e) in the SIR *algorithm can be formulated through an generalized eigenvalue problem as defined in (1.4) with* $\mathbf{M}_1 = \widehat{\mathbb{V}\mathrm{ar}}\left(\mathbb{E}(\mathbf{X} - \mathbb{E}(\mathbf{X})|Y)\right), \mathbf{M}_2 = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$*, i.e. we replace the population quantities in (1.35) by their estimates. Note also that applying the transformation given by (1.5) corresponds to standardizing* $\mathbf{X}$ *by setting* $\mathbf{Z}_i = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$ *in the* SIR *algorithm.*

1.10 Likelihood based sdr

The development of the sliced inverse regression (SIR) method and Theorem 3 gave rise to the likelihood based methods such as *principal fitted components* PFC [CF08] or *likelihood acquired directions* LAD [CF09], for a more detailed overview see [Coo07]. They are based on the the inverse regression $\mathbf{X}|Y$, but instead of assuming the linear design condition (LDC) given by (1.27), it is directly assumed that $\mathbf{X}|Y$ follows a parametric model, e.g. $\mathbf{X}|Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Delta}_y)$ with $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{\nu}_y$. If $\boldsymbol{\Delta}_y = \sigma^2\mathbf{I}_p$, then the errors are called isotropic.

Then the regression is fitted by maximum likelihood or other methods depending on the model assumptions, and the estimator for the *central subspace* is given by $\widehat{\mathcal{S}_{Y|\mathbf{X}}} = \mathrm{span}\{\widehat{\mathbf{B}}\}$ with $\widehat{\mathbf{B}}$ the maximum likelihood estimate for $\mathbf{B}$. Depending on the concrete assumptions placed on $\boldsymbol{\nu}_y$ and $\boldsymbol{\Delta}_y$ the resulting methods are:

- Principal components: Isotropic errors and no specific structure for $\boldsymbol{\nu}_y$.
- Isotropic PFC: Isotropic errors and $\boldsymbol{\nu}_y = \boldsymbol{\gamma}\left(\mathbf{f}_y - \mathbb{E}(\mathbf{f}_y)\right)$ with $\mathbf{f}_y = (f_1(y), \ldots, f_r(y))^T \in \mathbb{R}^r$ a known-vector valued function of y and $\boldsymbol{\nu} \in \mathbb{R}^{k\times r}$ a rank r matrix with unrestricted coefficients such that $k \leq \min(p, r)$.
- Structured PFC: $\boldsymbol{\nu}_y$ is modeled as above but the error structure has a linear structure and is independent of y, i.e. $\boldsymbol{\Delta}_y = \sum_{l=1}^{m} w_l\mathbf{M}_l$ with known matrizes $\mathbf{M}_l$.
- PFC: $\boldsymbol{\nu}_y$ modeled as above and with an error structure independent of y, i.e. $\boldsymbol{\Delta}_y = \boldsymbol{\Delta}$.
- LAD: A general structure for $\boldsymbol{\mu}_y$ and $\boldsymbol{\Delta}_y$ but requires a categorical response Y, for more details see [CF09].

1.11 Sliced Average Variance Estimation (save)

To overcome the disadvantage of non-exhaustiveness of SIR, [CW91] developed the so called *sliced average variance estimation* (SAVE). It is a second order inverse regression method

that is exhaustive, i.e. can recover the whole *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$, if the *linear design condition* (LDC), (1.27), and *constant conditional variance assumption* (CCV), (1.36), and some mild regularity are fulfilled. SAVE targets the second conditional moment, i.e. the conditional variance $\mathbb{V}\mathrm{ar}(\mathbf{X}|Y)$.

The *constant conditional variance assumption* (CCV) holds if and only if

$$\mathbb{V}\mathrm{ar}(\mathbf{X} \mid \mathbf{B}^T\mathbf{X}) \quad \text{is a nonrandom matrix for all} \quad \mathbf{B} \in \mathcal{S}(p,k) \tag{1.36}$$

i.e. constant in the conditioning argument.

Note that the multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ fulfills the *linear design condition* (LDC), (1.27), and *constant conditional variance assumption* (CCV), (1.36).

Theorem 5 gives the theoretical motivation of the SAVE method.

Theorem 5. *Assume model* (1.13)*, the linear design condition (LDC) in* (1.27)*, and the constant conditional variance condition (CCV) in* (1.36)*. Then*

$$\mathrm{span}\{\boldsymbol{\Sigma}_{\mathbf{x}} - \mathbb{V}\mathrm{ar}(\mathbf{X}|Y)\} \subseteq \boldsymbol{\Sigma}_{\mathbf{x}}\mathcal{S}_{Y|\mathbf{X}}$$

The proof can be found in [CW91] or in [Li18] Theorem 5.1. If Theorem 5 and Theorem 2 are applied with $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \mathbb{E}(\mathbf{X}))$ we get

$$\mathrm{span}\{\mathbf{I}_p - \mathbb{V}\mathrm{ar}(\mathbf{Z}|Y)\} \subseteq \mathcal{S}_{Y|\mathbf{Z}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{Y|\mathbf{X}} \tag{1.37}$$

Equation (1.37) will be used as the estimation equation of SAVE on the population level. The *sliced average variance estimation* (SAVE) algorithm is given by

Let $(Y_i, \mathbf{X}_i)_{i=1}^n$ be an i.i.d. sample from model (1.13), assume the linear design condition (LDC) given by (1.27) and the constant conditional variance condition in (1.36). Then the SAVE [CW91] algorithm is given by

(a) Standardize $\mathbf{Z}_i = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}(\mathbf{X}_i - \bar{\mathbf{X}})$ where $\bar{\mathbf{X}} = (1/n)\sum_{i=1}^n \mathbf{X}_i$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} = (1/n)\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$.

(b) Divide the range of Y into $H \in \mathbb{N}$ slices $I_h = (q_{h-1}, q_h]$ for $h = 1, \ldots, H$, discretize Y via $\tilde{Y} = \sum_{h=1}^H h 1_{\{Y \in \mathbf{I}_h\}}$, and set $n_h = \sum_{i=1}^n 1_{\{\tilde{Y}_i = h\}} = \#\{i : \tilde{Y}_i = h\} = \#\{i : Y_i \in I_h\}$.

(c) For each slice h calculate the sample conditional variance of $\mathbb{V}\mathrm{ar}(\mathbf{Z}|\tilde{Y} = h)$ by

$$\widehat{\mathbb{V}\mathrm{ar}}(\mathbf{Z}|\tilde{Y} = h) = \frac{\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T 1_{\{\tilde{Y}_i = h\}}}{\sum_{i=1}^n 1_{\{\tilde{Y}_i = h\}}} = \frac{1}{n_h}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i^T 1_{\{\tilde{Y}_i = h\}} = \frac{1}{n_h}\sum_{i:Y_i \in I_h} \mathbf{Z}_i\mathbf{Z}_i^T$$

(d) Calculate the matrix

$$\widehat{\mathbf{M}}_1 = \frac{1}{H}\sum_{h=1}^H n_h \left(\mathbf{I}_p - \widehat{\mathbb{V}\mathrm{ar}}(\mathbf{Z}|\tilde{Y} = h)\right)^2$$

(e) Calculate the spectral decomposition of $\widehat{\mathbf{M}}_1$ and set $\widehat{\mathbf{B}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}(\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_k)$ where $\hat{\mathbf{b}}_j$ for $j = 1, \ldots, k$ are the eigenvectors of $\widehat{\mathbf{M}}_1$ corresponding to the k largest eigenvalues

Remark. *Step (e) in the* SAVE *algorithm is exactly the same as in the* SIR *algorithm except that the matrix $\widehat{\mathbf{M}}_1$ is different, therefore it can be rewritten as an generalized eigenvalue problem (GEV) (see (1.4)) given by $(\widehat{\mathbf{M}}_1, \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}})$. As for* SIR *the dimension $k = \dim(\mathcal{S}_{Y|\mathbf{X}}) = rank(\mathbf{B})$ is assumed to be known in the algorithm but there are a number of asymptotic test and methods to estimate it. For further information see [CW91, Coo00, CY01, SCW07, BY11, Li18].*

1.12 Other sdr methods

So far the most prominent sufficient dimension reduction techniques based on inverse regression like SIR and SAVE have been discussed. We have seen that all of them can be interpreted as a generalized eigenvalue problem (GEV) $(\mathbf{M}_1, \mathbf{M}_2)$ with different matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ determining the methods on the population level, e.g. $\mathbf{M}_1 = \mathbb{V}\mathrm{ar}(\mathbb{E}(\mathbf{X}|Y))$ and $\mathbf{M}_2 = \mathbb{V}\mathrm{ar}(\mathbf{X})$ for SIR. Moreover different estimation methods for the corresponding population quantities can be deployed resulting in a number of different estimators for the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. There are a number of other methods based on inverse regression that can be casted into the generalized eigenvalue framework. Not all of them will be presented in such detail as SIR and SAVE but table 1.1 gives an overview of the different assumptions and matrizes $\mathbf{M}_1$ and $\mathbf{M}_2$ of the generalized eigenvalue problem that determines the methods. In table 1.1 the linear design condition given by (1.27) is abbreviated by LDC and the constant conditional variance condition given in (1.36) by CCV. Moreover CR stands for *contour regression* [LZC05b, Li18], DR for *directional regression* [LW07b, Li18]), and pHd for *principal Hessian direction* [Li92, Coo98c, Li18] which was further developed by [Coo98a, CL02, CL04]. Further in table 1.1 we set $\mathbf{Z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \mathbb{E}(X))$, the tilde overset represents an independent copy of the corresponding quantity, i.e. $\tilde{\mathbf{Z}}$ is an independent copy of $\mathbf{Z}$, and $\delta > 0$ denotes a small constant. Note that except for pHd all methods in table 1.1 are based on the inverse regression of $\mathbf{X}$ on Y (pHd is included for completeness since it can also be formulated as an generalized eigenvalue problem).

Table 1.1: SDR techniques as GEV $(\mathbf{M}_1, \mathbf{M}_2)$ as in (1.4)

Assumptions	Method	$\mathbf{M}_1$	$\mathbf{M}_2$
	PCA	$\boldsymbol{\Sigma}_{\mathbf{x}}$	$\mathbf{I}_p$
LDC	SIR	$\mathbb{V}\mathrm{ar}(\mathbb{E}(\mathbf{X} \mid Y))$	$\boldsymbol{\Sigma}_{\mathbf{x}}$
LDC	PFC	$\boldsymbol{\Sigma}_{\mathbb{E}(\mathbf{X}\|Y)}$	$\boldsymbol{\Sigma}_{\mathbf{x}}$
LDC, CCV	SAVE	$\boldsymbol{\Sigma}_{\mathbf{x}} - \mathbb{V}\mathrm{ar}(\mathbf{X}\|Y)$	$\boldsymbol{\Sigma}_{\mathbf{x}}$
LDC, CCV	CR	$\left(2\mathbf{I}_p - 2\mathbb{E}\left(\mathbf{Z}\mathbf{Z}^T \|\|Y - \tilde{Y}\| < \delta\right) - 2\mathbb{E}\left(\mathbf{Z}\tilde{\mathbf{Z}}^T \|\|Y - \tilde{Y}\| < \delta\right)\right)^2$	$\boldsymbol{\Sigma}_{\mathbf{x}}$
LDC, CCV	DR	$\mathbb{E}\left((\mathbf{Z} - \tilde{\mathbf{Z}})(\mathbf{Z} - \tilde{\mathbf{Z}})^T \|Y, \tilde{Y}\right) - 2\mathbf{I}_p$	$\boldsymbol{\Sigma}_{\mathbf{x}}$
LDC, CCV	pHd	$\mathbb{E}\left((Y - \mathbb{E}(Y))\mathbf{Z}\mathbf{Z}^T\right)$	$\boldsymbol{\Sigma}_{\mathbf{x}}$

1.13 Principal Hessian Direction (phd)

Principal Hessian direction (pHd) is a sufficient dimension reduction technique based on the forward regression developed by [Li92] and later refined by [Coo98b] and [CL02, CL04]. Assume model (1.18), i.e. $\mathbb{E}(Y \mid \mathbf{X}) = \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{X})$, then pHd is based on the observation given in (1.38).

$$H(\mathbf{x}) = \frac{\partial^2 \mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \frac{\partial^2 \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \mathbf{B} \frac{\partial^2 \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{X} = \mathbf{B}^T\mathbf{x})}{\partial (\mathbf{B}^T\mathbf{x}) \partial (\mathbf{x}^T\mathbf{B})} \mathbf{B}^T \quad (1.38)$$

Therefore we can conclude $\operatorname{span}\{H(\mathbf{x})\} \subseteq \operatorname{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$ and the same holds true for $\mathbb{E}(H(\mathbf{X}))$. Then set $\mathbf{Z} = \mathbf{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{X} - \mathbb{E}(\mathbf{X}))$ and under the assumption that $\mathbf{Z}$ is normal distributed Li showed using Stein's Lemma that

$$\operatorname{span}\{\mathbb{E}(H(\mathbf{X}))\} = \operatorname{span}\{\mathbb{E}\left((Y - \mathbb{E}(Y))\mathbf{Z}\mathbf{Z}^T\right)\} \subseteq \operatorname{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{Z}}$$

The pHd estimator is then defined as the first $k = \dim(\mathcal{S}_{Y|\mathbf{X}}) = \operatorname{rank}(\mathbf{B})$ eigenvectors of the generalized eigenvalue problem (GEV) given by $\mathbf{M}_1 = \mathbb{E}\left((Y - \mathbb{E}(Y))\mathbf{Z}\mathbf{Z}^T\right)$ and $\mathbf{M}_2 = \mathbf{\Sigma}_{\mathbf{x}}$ where all expectations are replaced by sample averages.

1.14 Minimum Average Variance Estimation (mave)

So far except from principal Hessian direction (pHd) all sufficient dimension reduction methods presented are based on the inverse regression given by Theorem 3. This Theorem assures the connection of the inverse regression to the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ under the linear design condition (LDC). Therefore if we want to get rid of the LDC the inverse regression cannot be used anymore. In [XTLZ02] the *mimum average variance estimator* (MAVE) for $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, which is based on the *forward regression* in (1.18) with $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$ is proposed.

The target function of *mimum average variance estimator* (MAVE) on the population level is given by

$$T(\mathbf{V}) = \mathbb{E}(Y - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X}))^2 = \mathbb{E}\left(\underbrace{\mathbb{E}[(Y - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X}))^2|\mathbf{V}^T\mathbf{X}]}_{=\sigma^2(\mathbf{V}^T\mathbf{X})}\right) \quad (1.39)$$

$$= \mathbb{E}(\sigma^2(\mathbf{V}^T\mathbf{X})) \quad (1.40)$$

for $\mathbf{V} \in \mathcal{S}(p, k)$.

Theorem 6. *Assume model (1.18) holds then we have*

$$\operatorname{span}\{\mathbf{B}\} := \operatorname{span}\{\operatorname{argmin}_{\mathbf{V} \in \mathcal{S}(p,k)} T(\mathbf{V})\} \quad (1.41)$$

for $T(\mathbf{V})$ defined in (1.40).

Proof of Theorem 6. From $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$ in (1.18) it follows

$$\begin{aligned} T(\mathbf{V}) &= \mathbb{E}\left(g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})\right)^2 + 2\mathbb{E}\left([g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})]\epsilon\right) + \mathbb{V}\mathrm{ar}(\epsilon) \\ &= \mathbb{E}\left(g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})\right)^2 + \mathbb{V}\mathrm{ar}(\epsilon) \geq \mathbb{V}\mathrm{ar}(\epsilon) \end{aligned} \tag{1.42}$$

since $\mathbb{E}\left([g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})]\epsilon\right) = \mathbb{E}\left([g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})]\underbrace{\mathbb{E}(\epsilon \mid \mathbf{X})}_{=0}\right) = 0$ due to $\mathbb{E}(\epsilon \mid \mathbf{X}) = 0$.

(1.42) and the discussion about the search space for sufficient dimension reduction in (1.2), yield

$$T(\mathbf{V}) = \mathbb{V}\mathrm{ar}(\epsilon) = \eta^2$$

for all $\mathbf{V}$ such that $\mathrm{span}\{\mathbf{V}\} = \mathrm{span}\{\mathbf{B}\}$. For all $\mathbf{V}$ such that $\mathrm{span}\{\mathbf{V}\} \neq \mathrm{span}\{\mathbf{B}\}$ it holds $\mathbb{E}(Y|\mathbf{V}^T\mathbf{X}) \neq g(\mathbf{B}^T\mathbf{X})$ and

$$T(\mathbf{V}) = \mathbb{E}\left(g(\mathbf{B}^T\mathbf{X}) - \mathbb{E}(Y|\mathbf{V}^T\mathbf{X})\right)^2 + \mathbb{V}\mathrm{ar}(\epsilon) > \mathbb{V}\mathrm{ar}(\epsilon) = \eta^2$$

This completes the proof. □

For the estimation of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathrm{span}\{\mathbf{B}\}$ from an i.i.d. sample $(Y_i, \mathbf{X}_i)_{i=1}^n$ of model (1.18) we replace the target function $T(\mathbf{V})$ in Theorem 6 by an estimate $T_n(\mathbf{V})$, where $\mathbf{V} \in \mathcal{S}(p,k)$. A local linear expansion of $\mathbb{E}(Y_i \mid \mathbf{V}^T\mathbf{X}_i)$ around $\mathbf{X}_0$ yields

$$\mathbb{E}(Y_i \mid \mathbf{V}^T\mathbf{X}_i) \approx a + \mathbf{b}^T\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_0), \tag{1.43}$$

where $a = g(\mathbf{V}^T\mathbf{X}_0) \in \mathbb{R}$, $\mathbf{b} = \nabla g(\mathbf{V}^T\mathbf{X}_0) \in \mathbb{R}^k$. Therefore we obtain the following approximation for $\sigma^2(\mathbf{V}^T\mathbf{X}_0)$ in (1.40),

$$\sigma^2(\mathbf{V}^T\mathbf{X}_0) \approx \sum_{i=1}^n (Y_i - \mathbb{E}(Y \mid \mathbf{V}^T\mathbf{X}))^2 w_{i,0} \approx \sum_{i=1}^n (Y_i - a - \mathbf{b}^T\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_0))^2 w_{i,0}$$

for some weights $w_{i,0}$ that sum to 1 ($\sum_i w_{i,0} = 1$). The weights play a crucial role in the estimation. They are given by

$$w_{i,0}(\mathbf{V}) := \frac{K\left(\frac{\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_0)}{h}\right)}{\sum_l K\left(\frac{\mathbf{V}^T(\mathbf{X}_l - \mathbf{X}_0)}{h}\right)}, \tag{1.44}$$

for a k dimensional kernel $K(\cdot)$, and a bandwidth $h \in \mathbb{R}_+$.

It is common to set $K(\cdot) = \tilde{K}(\|\cdot\|_2)$ for a monotone decreasing univariate kernel $\tilde{K}(\cdot) : \mathbb{R} \to \mathbb{R}_+$, since [XTLZ02] suggested that the weights depend only on the distance of $\mathbf{V}^T(\mathbf{X}_l - \mathbf{X}_0)$ (the further $\mathbf{V}^T\mathbf{X}_i$ is away from $\mathbf{V}^T\mathbf{X}_0$ the worse the linear expansion is and the less weight we assign).

Then, an estimator for $\sigma^2(\mathbf{V}^T\mathbf{X}_0)$ in (1.40) is given by

$$\hat{\sigma}^2(\mathbf{V}^T\mathbf{X}_0) := \min_{a,\mathbf{b}} \sum_{i=1}^{n} (Y_i - a - \mathbf{b}^T\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_0))^2 w_{i,0}(\mathbf{V}) \tag{1.45}$$

and the target function $T(\mathbf{V})$ is estimated by

$$T_n(\mathbf{V}) = \frac{1}{n}\sum_{j=1}^{n} \hat{\sigma}^2(\mathbf{V}^T\mathbf{X}_j) = \frac{1}{n} \min_{a_j,\mathbf{b}_j : j=1,..,n} \sum_j \sum_i (Y_i - a_j - \mathbf{b}_j{}^T\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_j))^2 w_{i,j}(\mathbf{V}) \tag{1.46}$$

where the weights are given in (1.44). The Gaussian kernel is usually used with bandwidth satisfying $h = h_n \propto n^{-1/(4+k)}$, as typically done in nonparametric function estimation, in order to obtain optimal asymptotic properties.

Then $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ is estimated by

$$\widehat{\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}} := \text{span}\{\text{argmin}_{\mathbf{V}\mathcal{S}(p,k)}\, T_n(\mathbf{V})\}.$$

1.15 Outer product gradient (opg)

The OPG estimator [XTLZ02], introduced in the same paper as MAVE, is conceptually similar to pHd and MAVE. The theoretical motivation is given by

Theorem 7. *Assume model (1.18) and let*

$$\mathbf{\Sigma}_\nabla = \mathbb{E}\left(\nabla g(\mathbf{B}^T\mathbf{X})\nabla g(\mathbf{B}^T\mathbf{X})^T\right)$$

where $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{x}) = g(\mathbf{B}^T\mathbf{x})$. *Then* $\text{span}\{\mathbf{\Sigma}_\nabla\} = \text{span}\{\mathbf{B}\}$

Proof of Theorem 7. From model (1.18) we have $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}(Y \mid \mathbf{B}^T\mathbf{x}) = g(\mathbf{B}^T\mathbf{x})$ and taking the derivative, yields

$$\nabla_{\mathbf{x}} g(\mathbf{B}^T\mathbf{x}) = \nabla_{\mathbf{B}^T\mathbf{x}} g(\mathbf{B}^T\mathbf{x}) \frac{\partial(\mathbf{B}^T\mathbf{x})}{\partial \mathbf{x}} = \nabla g(\mathbf{B}^T\mathbf{x})\mathbf{B}$$

and therefore we conclude

$$\mathbf{\Sigma}_\nabla = \mathbb{E}\left(\nabla_{\mathbf{x}} g(\mathbf{B}^T\mathbf{X})\nabla_{\mathbf{x}} g(\mathbf{B}^T\mathbf{X})^T\right) = \mathbf{B}\mathbb{E}\left(\nabla g(\mathbf{B}^T\mathbf{x})\nabla g(\mathbf{B}^T\mathbf{x})^T\right)\mathbf{B}^T$$

which completes the proof. □

For $\mathbf{V} = \mathbf{I}_p$ in (1.46) calculate $(a_j, \mathbf{b}_j^T)_{j=1}^n$ by solving the optimization given in (1.46). Then $\mathbf{b}_j$ is an estimate for $\nabla g(\mathbf{B}^T\mathbf{X}_j)$ and set

$$\widehat{\mathbf{\Sigma}}_\nabla := \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}_j\mathbf{b}_j^T \tag{1.47}$$

as an estimator for Σ_∇ in Theorem 7.

Then the Outer product gradient (OPG) estimator for $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ is defined as the span of the first k eigenvectors of $\widehat{\mathbf{\Sigma}}_\nabla$.

Remark. *The* OPG *estimator can be used as initial value for the optimization procedure of the* MAVE.

1.16 Contributions of the thesis

The contributions of this thesis are described next. The methodology and results formed the basis for three papers of the author, [FB21a, FB21b, KFB21]. All three papers were supervised by Prof. Efstathia Bura. The third paper is joint work with Daniel Kapla, a PhD colleague, with equally divided contributions.

First the novel *conditional variance estimator* CVE is presented in Chapter 2. CVE is a sufficient dimension reduction method that is consistent for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ in model (1.18) under weak regularity assumptions. As MAVE, it differentiates itself from most other inverse regression based SDR methods, as it is based on the forward model $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$ with an additive error term in (1.18). The *conditional variance estimator* CVE differs from other approaches, including MAVE, in that it only targets the $\text{span}\{\mathbf{B}\}$ and does not require an explicit form or estimation of the link function g. As a result, it requires weaker assumptions on its smoothness. Further, the accuracy and performance of the novel CVE estimator is demonstrated via simulations and in data applications, which indicate that CVE is mostly on par or better than MAVE, the gold standard in sufficient dimension reduction so far.

In Chapter 3, the novel NN − SDR estimator is presented. Most forward regression SDR methods that exhibit excellent estimation performance, like MAVE and CVE, are usable for relatively small predictor dimensions, p, and sample sizes, n. When both p and n increase substantially, their computation can spread over days or weeks, thus rendering them infeasible in practice. The NN − SDR estimator combines forward regression SDR with neural networks in order to remove the limitation of small p and n. NN − SDR is a two stage estimator that carries out simultaneous sufficient dimension reduction and neural network learning of the link function in model (1.18). First we fit an arbitrary neural net to the data, and in the second stage we refine the estimate with a specific architecture using a bottleneck. The premise of the two stage NN−SDR estimator is conceptually similar to MAVE with the difference that we use neural nets as universal function approximators instead of nonparametric local linear smoothing methods. The advantage of this approach is that it retains the accuracy of state of the art SDR methods while it can be easily deployed to large scale regressions frequently encountered in applications as is demonstrated via simulations and data examples. It also obtains predictions at nearly no additional computational cost compared to fully non-parametric methods used for predictions in MAVE and CVE. Further, the extension of the proposed NN − SDR estimator to online learning, where new data are dynamically added, is straightforward.

In Chapter 4, the novel *ensemble conditional variance estimator* ECVE is presented. ECVE uses the idea of ensembles, i.e. families of functions that are rich enough in a certain sense, to extend the CVE for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ to the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. ECVE is shown to be a consistent estimator for $\text{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$ in model (1.13). Moreover, the performance of ECVE is compared via simulations and data examples against the state of the art estimation method, *central subspace mean average variance estimation* CSMAVE, for

the central subspace $\mathcal{S}_{Y|\mathbf{X}}$. The accuracy of ECVE is mostly on par or better than CSMAVE, which is the extension of MAVE.

2 Conditional Variance Estimation for the mean subspace

In this chapter we introduce the novel *conditional variance estimator* for estimating the *mean subspace* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, see Section 1.5, for the model (1.18). The estimator is based on a new approach or target function that is, to the best knowledge of the author, never seen before in the literature.

Throughout the chapter we refer to the following assumptions as needed.

(A.1). *Model* $Y = g(\mathbf{B}^T\mathbf{X})+\epsilon$ *holds with* $Y \in \mathbb{R}$, $g : \mathbb{R}^k \to \mathbb{R}$ *non constant in all arguments,* $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_k) \in \mathbb{R}^{p\times k}$ *of rank* $k \leq p$, $\mathbf{X} \in \mathbb{R}^p$ *independent from* ϵ, *the distribution of* $\mathbf{X}$ *is absolute continuous with respect to the Lebesgue measure in* $\mathbb{R}^p$, *the support of the density* $f_{\mathbf{X}}$ *is convex,* $\mathbb{V}\mathrm{ar}(\mathbf{X}) = \mathbf{\Sigma}_{\mathbf{x}}$ *is positive definite,* $\mathbb{E}(\epsilon) = 0$, $\mathbb{V}\mathrm{ar}(\epsilon) = \eta^2 < \infty$.

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

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

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

Remark. *The mean subspace and the central subspace agree in this model, i.e.* $\mathrm{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \mathcal{S}_{Y|\mathbf{X}}$ *due to* $\mathbf{X}$ *independent to* ϵ. *Assumption (A.1) guarantees the existence and uniqueness of* $\mathrm{span}\{\mathbf{B}\}$, *i.e. the mean subspace, by proposition 6.4 in [Coo98c]. Assumption (A.4) is not as restrictive as it might seem. In Proposition 11 of [YLC08] it is shown that there is a compact set* $\mathcal{S} \subset \mathbb{R}^p$ *such that the mean subspace of model (1.18) is the same as the mean subspace of* $Y = g(\mathbf{B}^T\mathbf{X}_{|\mathcal{S}}) + \epsilon$, *where* $\mathbf{X}_{|\mathcal{S}} = \mathbf{X}1_{\{\mathbf{X}\in\mathcal{S}\}}$ *and* 1_A *is the indicator function of* A. *Further* $\mathcal{S}$ *can be assumed to be an ellipsoid and for all* $\widetilde{\mathcal{S}} \supseteq \mathcal{S}$ *the same assertion holds true.*

2.1 Motivation and Definitions

Definition 9. *For an integer* q *with* $q \leq p$ *and any* $\mathbf{V} \in \mathcal{S}(p, q)$, *we define*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mathbb{V}\mathrm{ar}(Y \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}), \tag{2.1}$$

where $\mathbf{s}_0 \in supp(f_{\mathbf{X}})$ *is a shifting point.*

Definition 10. *For* $\mathbf{V} \in \mathcal{S}(p, q)$, *we define the objective function,*

$$L(\mathbf{V}) = \int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right). \tag{2.2}$$

$L(\mathbf{V})$ in (2.2) is the objective function on the population level for the estimator we propose for the span of $\mathbf{B}$ in (1.18) and Theorem 9 provides the statistical motivation for the objective function (2.2) of the conditional variance estimator. First in Theorem 8 we derive that both population based functions (2.1) and (2.2) are well defined.

Theorem 8. *Let $\mathbf{X}$ be a p-dimensional continuous random vector with density $f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{s}_0 \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$, and $\mathbf{V} \in S(p,q)$ defined in (1.1).*

(a) Under assumption (A.2), it holds that

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{r}_1) = \frac{f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r})d\mathbf{r}} \tag{2.3}$$

is a proper density of $\mathbf{X}$ conditioned on $\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ that is concentrated on the affine subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$.

(b) Under assumptions (A.1), (A.2) and (A.4) it holds that $\tilde{L}(\mathbf{V},\mathbf{s}_0)$ in (2.1) and $L(\mathbf{V})$ in (2.2) are well defined and continuous. Moreover, it holds

$$\tilde{L}(\mathbf{V},\mathbf{s}_0) = \mu_2(\mathbf{V},\mathbf{s}_0) - \mu_1(\mathbf{V},\mathbf{s}_0)^2 + \eta^2 \tag{2.4}$$

where

$$\mu_l(\mathbf{V},\mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1)^l \frac{f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r})d\mathbf{r}} d\mathbf{r}_1 = \frac{t^{(l)}(\mathbf{V},\mathbf{s}_0)}{t^{(0)}(\mathbf{V},\mathbf{s}_0)} \tag{2.5}$$

with

$$t^{(l)}(\mathbf{V},\mathbf{s}_0) = \int_{\mathbb{R}^q} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1)^l f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1. \tag{2.6}$$

Proof of Theorem 8. For part (a), note 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 guarantees the existence of the regular conditional probability of $\mathbf{X} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ by Theorem 3.1 of [LJFR04], see also [Fad85]. Further the measure is concentrated on $\mathbf{s}_0 + \text{span}\{\mathbf{V}\} \cap \text{supp}(f_{\mathbf{X}}) \subset \mathbb{R}^p$ and is given by (2.3) due to the orthogonal decomposition (1.3) and Definition 8.38 and Theorem 8.39 of [Kar93].

Furthermore, we can calculate explicitly:

$$\mathbb{P}(\mathbf{X} \le \mathbf{x} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) = \lim_{h\downarrow 0} \frac{\text{pr}(\{\mathbf{X} \le \mathbf{x}\} \cap \{\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}\})}{\text{pr}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\})} \tag{2.7}$$

where $\text{span}_h\{\mathbf{V}\} = \{\mathbf{x} \in \mathbb{R}^p : d(\mathbf{V},0) = \|\mathbf{x} - \mathbf{P}_{\mathbf{V}}\mathbf{x}\|^2 \le h\}$. Using the orthogonal decomposition (1.3) and writing $\mathbf{W} = (\mathbf{V},\mathbf{U}) \in S(p,p)$ (i.e. $\mathbf{W}\mathbf{W}^T = \mathbf{I}_p$). Inserting $\mathbf{x} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{x}-\mathbf{s}_0) + \mathbf{U}\mathbf{U}^T(\mathbf{x}-\mathbf{s}_0) = \mathbf{s}_0 + \mathbf{W}\mathbf{W}^T(\mathbf{x}-\mathbf{s}_0)$ to obtain the second equality below

$$\begin{aligned}\mathbb{P}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}) &= \int_{\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\ &= \int_{\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{W}\underbrace{\mathbf{W}^T(\mathbf{x}-\mathbf{s}_0)}_{=\mathbf{r}=(\mathbf{r}_1,\mathbf{r}_2)^T\in\mathbb{R}^p})d\mathbf{x} =\end{aligned} \tag{2.8}$$

with $d\mathbf{r} = d\mathbf{x}$ due to the integral transformation rule (i.e. the Jacobi determinant of shifting and multiplying with an orthogonal matrix is 1). Further note that

$$\mathbf{x} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\} \iff h \geq d(\mathbf{V}, \mathbf{s}_0) = \|(\mathbf{x}-\mathbf{s}_0) - \underbrace{\mathbf{P}_\mathbf{V}}_{=\mathbf{V}\mathbf{V}^T}(\mathbf{x}-\mathbf{s}_0)\| \tag{2.9}$$
$$= \|\mathbf{V}\mathbf{V}^T(\mathbf{x}-\mathbf{s}_0) + \mathbf{U}\mathbf{U}^T(\mathbf{x}-\mathbf{s}_0) - \mathbf{V}\mathbf{V}^T\left(\mathbf{V}\mathbf{V}^T(\mathbf{x}-\mathbf{s}_0) + \mathbf{U}\mathbf{U}^T(\mathbf{x}-\mathbf{s}_0)\right)\|^2$$
$$= \|\mathbf{U}\underbrace{\mathbf{U}^T(\mathbf{x}-\mathbf{s}_0)}_{\mathbf{r}_2}\|^2 = \|\mathbf{r}_2\|^2$$

due to $\mathbf{V}^T\mathbf{U} = \mathbf{0}$ and the fact that multiplying with an orthonormal matrix $\mathbf{U}$ does not change the norm. Inserting (2.9) in (2.8) obtains

$$\mathbb{P}(\mathbf{X} \in \mathbf{s}_0 + \text{span}_h\{\mathbf{V}\}) = \int_{\{\mathbf{r}_1\in\mathbb{R}^q, \mathbf{r}_2\in\mathbb{R}^{p-q}\}\cap\{\|\mathbf{r}_2\|^2\leq h\}} f_\mathbf{X}(\mathbf{s}_0 + W\mathbf{r})d\mathbf{r} \tag{2.10}$$
$$= \int_{\mathbb{R}^q}\int_{\|\mathbf{r}_2\|^2\leq h} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1$$
$$= h^{(p-q)/2}\int_{\mathbb{R}^q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1$$

where the second equality is due to Fubini's Theorem and the last equality follows due to the substitution $\mathbf{r}_2/h^{1/2} = \tilde{\mathbf{r}}_2$ and $\{\|\mathbf{r}_2\|^2 \leq h\} = \{\|\mathbf{r}_2/h^{1/2}\|^2 \leq 1\} = \{\|\tilde{\mathbf{r}}_2\|^2 \leq 1\}$.

With the same calculations the numerator of (2.7) equals

$$\text{pr}(\{\mathbf{X}\leq\mathbf{x}\}\cap\{\mathbf{X}\in\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}\}) = \int_{\{\mathbf{z}\leq\mathbf{x}\}\cap\{\mathbf{z}\in\mathbf{s}_0+\text{span}_h\{\mathbf{V}\}\}} f_\mathbf{X}(\mathbf{z})d\mathbf{z} \tag{2.11}$$
$$= \int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q}\int_{\|\mathbf{r}_2\|^2\leq h} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2)d\mathbf{r}_2 d\mathbf{r}_1$$
$$= h^{(p-q)/2}\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1$$

where $(y_1,\dots,y_q)^T = \mathbf{V}^T(\mathbf{x}-\mathbf{s}_0)$. Observe that if $\mathbf{x}-\mathbf{s}_0$ ranges in the orthogonal complement of $\text{span}\{\mathbf{V}\}$, i.e. $(y_1,\dots,y_q)^T = 0$ by (2.9) and therefore the cdf is constant (i.e. the density is concentrated on $\mathbf{s}_0 + \text{span}\{\mathbf{V}\} \cap \text{supp}(f_\mathbf{X})$). Substituting the numerator (2.11) and denominator (2.10) into (2.7) yields

$$\lim_{h\downarrow 0}\frac{\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1}{\int_{\mathbb{R}^q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1 + h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1} \tag{2.12}$$

A quotient converges if numerator and denominator converge and the denominator converges to a non zero constant. Therefore we will argue by the dominated convergence Theorem that this is the case. Due to (A.4) (i.e. $\text{supp}(f_\mathbf{X})$ is compact) all integrals are over compact sets (i.e. this is suppressed in the notation). Due to (A.2) the integrand is continuous and therefore can be bounded by $\sup_{\mathbf{x}\in\text{supp}(f_\mathbf{X})} f_\mathbf{X}(\mathbf{x}) < \infty$ (i.e. a continuous function attains a finite maximum on a compact set). Since the integrals are over compact

sets $\sup_{\mathbf{x}\in\text{supp}(f_{\mathbf{X}})} f_{\mathbf{X}}(\mathbf{x})$ is an integrable dominating function and the dominated convergence Theorem can be applied to pass the limit $\lim_{h\downarrow 0}$ under the integral and due to the continuity of $f_{\mathbf{X}}$ also inside the function argument, i.e. for the denominator we get

$$\begin{aligned}
&\lim_{h\downarrow 0}\int_{\mathbb{R}^q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1 \qquad (2.13)\\
&=\int_{\mathbb{R}^q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+\underbrace{\lim_{h\downarrow 0} h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2}_{=0})d\tilde{\mathbf{r}}_2 d\mathbf{r}_1\\
&=\underbrace{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1}_{=t^{(0)}(\mathbf{V},\mathbf{s}_0)}\left(\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} d\tilde{\mathbf{r}}_2\right)>0
\end{aligned}$$

Further note that $t^{(0)}(\mathbf{V},\mathbf{s}_0)>0$ due to continuity of $f_{\mathbf{X}}$, $\mathbf{s}_0\in\text{supp}(f_{\mathbf{X}})$ (which is meant as $\mathbf{s}_0$ being in the interior of the support), and the support being convex (i.e. in a small neighbourhood around $\mathbf{s}_0$ the density is strictly positive due to continuity).

With exactly the same reasoning we obtain for the numerator

$$\begin{aligned}
&\lim_{h\downarrow 0}\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q}\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+h^{1/2}\mathbf{U}\tilde{\mathbf{r}}_2)d\tilde{\mathbf{r}}_2 d\mathbf{r}_1 \qquad (2.14)\\
&=\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1\left(\int_{\|\tilde{\mathbf{r}}_2\|^2\leq 1} d\tilde{\mathbf{r}}_2\right)
\end{aligned}$$

Finally plugging in (2.13) and (2.14) into (2.12) yields

$$\mathbb{P}(\mathbf{X}\leq\mathbf{x}\mid\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\})=\frac{\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1} \qquad (2.15)$$

where $(y_1,\dots,y_q)^T=\mathbf{V}^T(\mathbf{x}-\mathbf{s}_0)$. Note that $f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{x})$ defined in (2.3) full fills

$$\mathbb{P}(\mathbf{X}\leq\mathbf{x}\mid\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\})=\int_{-\infty}^{y_1}\dots\int_{-\infty}^{y_q} f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(\mathbf{r})drs$$

and thus is the density of $\mathbf{X}\mid\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}$. This completes the proof of part (a).

Next we show part (b) of the Theorem. Due to the independence of $\mathbf{X}$ and ϵ in (1.18), $\mathbb{V}\text{ar}(Y\mid\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\})=\mathbb{V}\text{ar}(g(\mathbf{B}^T\mathbf{X})\mid\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\})+\mathbb{V}\text{ar}(\epsilon)$. Using (2.3) and $\mathbb{V}\text{ar}(Y\mid Z)=\mathbb{E}(Y^2\mid Z)-\mathbb{E}(Y\mid Z)^2$, we obtain (2.4).

The parameter integral [Heu95],

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

is well defined and continuous if (1) $\tilde{g}(\mathbf{V},\mathbf{s}_0,\cdot)$ is integrable for all $\mathbf{V}\in S(p,q),\mathbf{s}_0\in\text{supp}(f_{\mathbf{X}})$, (2) $\tilde{g}(\cdot,\cdot,\mathbf{r})$ is continuous for all $\mathbf{r}$, and (3) there exists an integrable dominating function of $\tilde{g}$ that does not depend on $\mathbf{V}$ and $\mathbf{s}_0$ [see [Heu95, p. 101]].

Furthermore $t^{(l)}(\mathbf{V}, \mathbf{s}_0) = \int_K \tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r}) d\mathbf{r}$ for some compact set K since $\text{supp}(f_{\mathbf{X}})$ is compact due to (A.4). The function $\tilde{g}(\mathbf{V}, \mathbf{s}_0, \mathbf{r})$ is continuous in all inputs by the continuity of g and $f_{\mathbf{X}}$ due to (A.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.

Next $\mu_l(\mathbf{V}, \mathbf{s}_0) = t^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ is continuous since $t^{(0)}(\mathbf{V}, \mathbf{s}_0) > 0$ for all $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ by the analogue argument as in part (a). Then, $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (2.4) is continuous. Since $L(\mathbf{V})$ is a parameter integral, it is well defined and continuous following the same arguments as above. □

Theorem 8 (a) establishes that (2.3) is a proper density. Since $\mathbf{X}$ has a continuous distribution, the set $\{\omega \in \Omega : \mathbf{X}(\omega) \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\}$ has probability 0 if $q < p$, but Theorem 8 (b) shows that $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ in (2.1) and $L(\mathbf{V})$ in (2.2), are well-defined using the concept of regular conditional probability [LJFR04] and can be expressed by using (2.3).

Next, Theorem 9 provides the motivation for the objective function given in (2.2) and why it can be used to identify $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ in model (1.18)

Theorem 9. *Suppose* $\mathbf{V} = (\mathbf{v}_1, ..., \mathbf{v}_q) \in \mathcal{S}(p, q)$ *and* $q \in \{1, \dots, p\}$. *Under assumptions (A.1), (A.2) and (A.4),*

(a) For all $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ *and* $\mathbf{V}$ *such that there exist* $u \in \{1, ..., q\}$ *with* $\mathbf{v}_u \in \text{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \mathbb{V}\text{ar}(\epsilon) = \eta^2$ *and* $L(\mathbf{V}) > \eta^2$.

(b) For all $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$ *and* $\text{span}\{\mathbf{V}\} \perp \text{span}\{\mathbf{B}\}$, $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \eta^2$ *and* $L(\mathbf{V}) = \eta^2$.

Proof of Theorem 9. Let $\mathbf{s}_0 \in \text{supp}(f_{\mathbf{X}})$.By (2.1) and the independence of $\mathbf{X}$ and ϵ we obtain $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mathbb{V}\text{ar}\left(g(\mathbf{B}^T\mathbf{X}) + \epsilon \mid \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) = \mathbb{V}\text{ar}\left(g(\mathbf{B}^T\mathbf{X}) \mid \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) + \mathbb{V}\text{ar}(\epsilon)$. Moreover $\mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\} \iff \mathbf{X} = \mathbf{s}_0 + \mathbf{P}_{\mathbf{V}}(\mathbf{X} - \mathbf{s}_0)$ yields

$$\begin{aligned}\tilde{L}(\mathbf{V}, \mathbf{s}_0) &= \mathbb{V}\text{ar}\left(g(\mathbf{B}^T\mathbf{X}) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)\right) + \mathbb{V}\text{ar}(\epsilon) \\ &= \mathbb{V}\text{ar}\left(g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)) \mid \mathbf{X} = \mathbf{s}_0 + \mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)\right) + \eta^2 \geq \eta^2\end{aligned} \tag{2.16}$$

If $\mathbf{B}^T\mathbf{V} \neq 0$ we can conclude that $\mathbb{V}\text{ar}(\mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0)) = \mathbf{B}^T\mathbf{V}\mathbf{V}^T\mathbb{V}\text{ar}(\mathbf{X})\mathbf{V}\mathbf{V}^T\mathbf{B}$ has at least one component with positive variance since $\mathbb{V}\text{ar}(\mathbf{X}) > 0$. Since g is nonconstant in every argument $g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0))$ in (2.16) has strictly positive variance, therefore if $\mathbf{B}^T\mathbf{V} \neq 0$, it holds $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \eta^2$. If $\mathbf{B}^T\mathbf{V} = 0$, it holds $\mathbf{B}^T\mathbf{V}\mathbf{V}^T(\mathbf{X} - \mathbf{s}_0) = 0$ and therefore $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \eta^2$ since the first term in (2.16) is the variance of a constant and therefore 0. Since $\mathbf{s}_0$ is arbitrary yet constant, the statements for $L(\mathbf{V})$ follow.

Alternatively one could use (2.4) and Jensen's inequality to proof the statement of Theorem 9. Observe that by (2.4) we have

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 + \eta^2 \geq \eta^2 \tag{2.17}$$

since $\mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 \geq 0$ due to Jensen inequality. To see this, write $\mathbf{R} \sim \tilde{f}_{\mathbf{R}}(\mathbf{r}_1) = f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)/\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r})d\mathbf{r}$ and $Z = g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{R})$, then by Jensen $\mu_2(\mathbf{V}, \mathbf{s}_0) = \mathbb{E}(Z^2) \geq \mathbb{E}(Z)^2 = \mu_1(\mathbf{V}, \mathbf{s}_0)^2$

If $\mathbf{B}^T\mathbf{V} = 0$ we have

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

and it follows $\mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 = g(\mathbf{B}^T\mathbf{s}_0)^2 - g(\mathbf{B}^T\mathbf{s}_0)^2 = 0$, therefore $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \eta^2$. Then note that equality in Jensen's inequality for a strictly convex function is only achieved if the integrand $Z = g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{R})^l$ is a constant random variable. If $\mathbf{B}^T\mathbf{V} \neq 0$ and since $g(\cdot)$ is non constant in every argument and the density $f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)/\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r})d\mathbf{r}$ is non degenerate, we have Z a nondegenerate random variable and Jensen inequality holds strictly, i.e. $\mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 > 0$. Therefore if $\mathbf{B}^T\mathbf{V} \neq 0$ we have $\tilde{L}(\mathbf{V}, \mathbf{s}_0) > \eta^2$. Since $\mathbf{s}_0$ is arbitrary yet constant and the monotony of the expectation operator, the statements for $L(\mathbf{V})$ follow. □

Theorem 9 also has an intuitive geometrical interpretation for the proposed method. If $\mathbf{X}$ is not random, the deterministic function $Y = g(\mathbf{B}^T\mathbf{X})$ is constant in all directions orthogonal to $\mathbf{B}$ and varies in all other directions. If randomness is introduced, as in model (1.18), then the variation in Y stems only from ϵ in all directions orthogonal to $\mathbf{B}$. In all other directions the variation comprizes of the sum of the variation of ϵ and of $g(\mathbf{B}^T\mathbf{X})$. In consequence, the objective function (2.2) captures the variation of Y as $\mathbf{X}$ varies in the column space of $\mathbf{V}$ and is minimized in the directions orthogonal to $\mathbf{B}$. A more thorough intuitive explanation is given in Section 2.3 via an toy example.

We have shown that the objective function $L(\mathbf{V})$ in (2.2) is well defined and continuous in Section 2.1. Let

$$\mathbf{V}_q = \operatorname{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L(\mathbf{V}). \tag{2.18}$$

$\mathbf{V}_q$ is well defined as the minimizer of a continuous function over the compact set $\mathcal{S}(p, q)$. Nevertheless, $\mathbf{V}_q$ is not unique since for all orthogonal $\mathbf{O} \in \mathbb{R}^{q\times q}$ such that $\mathbf{O}\mathbf{O}^T = \mathbf{I}_q$, $L(\mathbf{V}\mathbf{O}) = L(\mathbf{V})$ as $L(\mathbf{V})$ depends on $\mathbf{V}$ only through span$\{\mathbf{V}\}$. Therefore (2.18) is unique as an optimization over the Grassmann manifold (1.2) by the uniquness of span$\{\mathbf{B}\}$ given by Assumption (A.1) and proposition 6.4 in [Coo98c].

Further, we can view (2.2) also as a function from the Grassmann manifold to $[0, \mathbb{R})$. To see this, suppose $\mathbf{V} \in \mathcal{S}(p, q)$ is an arbitrary basis of a subspace $\mathbf{M} \in Gr(p, q)$. We can identify $\mathbf{M}$ through the projection $\mathbf{P}_{\mathbf{M}} = \mathbf{V}\mathbf{V}^T$.

Let again $\mathbf{W} = (\mathbf{V}, \mathbf{U}) \in \mathcal{S}(p, p)$ given by the orthogonal decomposition given in (1.3)

(i.e. $\mathbf{x} = \mathbf{W}\mathbf{W}^T\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$), then by the Fubini-Tornelli Theorem we obtain

$$\tilde{t}^{(l)}(\mathbf{P_M}, \mathbf{s}_0) = \int_{\text{supp}(f_\mathbf{X})} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{P_M}\mathbf{x})^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{P_M}\mathbf{x})d\mathbf{x} \tag{2.19}$$

$$= \int_{\text{supp}(f_\mathbf{X})} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{P_M}\mathbf{W}\underbrace{\mathbf{W}^T\mathbf{x}}_{=(\mathbf{r}_1,\mathbf{r}_2)^T})^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{P_M}\mathbf{W}\underbrace{\mathbf{W}^T\mathbf{x}}_{=(\mathbf{r}_1,\mathbf{r}_2)^T})d\mathbf{x}$$

$$= \int_{\text{supp}(f_\mathbf{X})\cap\mathbb{R}^q}\int_{\text{supp}(f_\mathbf{X})\cap\mathbb{R}^{p-q}} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{P_M}(\mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2))^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{P_M}(\mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2))d\mathbf{r}_2 d\mathbf{r}_1$$

$$= \underbrace{\int_{\text{supp}(f_\mathbf{X})\cap\mathbb{R}^{p-q}} g(\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\mathbf{r}_1)^l f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1}_{=t^{(l)}(\mathbf{V},\mathbf{s}_0)}\left(\int_{\text{supp}(f_\mathbf{X})\cap\mathbb{R}^{p-q}} d\mathbf{r}_2\right)$$

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

Therefore $\tilde{t}^{(l)}(\mathbf{P_M}, \mathbf{s}_0)/\tilde{t}^{(0)}(\mathbf{P_M}, \mathbf{s}_0) = t^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ and $\mu_l(\cdot, \mathbf{s}_0)$ in (2.5) can also be viewed as a function from $Gr(p, q)$ to $\mathbb{R}$. If the optimization (2.18) is over $Gr(p, q)$, the objective function (2.2) has a unique minimum at $\text{span}\{\mathbf{B}\}^\perp$ by Theorem 9. Therefore $\mathbf{B}$ is not uniquely identifiable but its $\text{span}\{\mathbf{B}\}$ is.

Corollary 10 follows directly from Theorem 9 and provides the means for identifying the linear projections of the predictors satisfying (1.18).

Corollary 10. *Under the assumptions (A.1), (A.2), and (A.3) the solution of the optimisation problem* $\mathbf{V}_q$ *in* (2.18) *is well defined. Let* $k = \dim(\text{span}\{\mathbf{B}\})$ *and* $q = p - k$,

(a) $\text{span}\{\mathbf{V}_q\} = \text{span}\{\mathbf{B}\}^\perp$

(b) $\text{span}\{\mathbf{V}_q\}^\perp = \text{span}\{\mathbf{B}\}$

We next define the estimation equation on the population level for the sufficient reduction space, $\text{span}\{\mathbf{B}\}$, in (1.18), which is motivated by Theorem 9 and Corollary 10 (b).

Definition 11. *The estimation equation of the* ***Conditional Variance Estimator*** CVE *on the population level is given by any basis* $\mathbf{B}_{p-q}$ *of* $\text{span}\{\mathbf{V}_q\}^\perp$. *That is, the* CVE *of* $\mathbf{B}$ *is any* $\mathbf{B}_{p-q}$ *such that*

$$\text{span}\{\mathbf{B}_{p-q}\} = \text{span}\{\mathbf{V}_q\}^\perp \tag{2.20}$$

When $q = p - k$, where $k = \text{rank}(\mathbf{B})$ in (1.18), then the CVE obtains the population $\text{span}\{\mathbf{B}\}$. Alternatively, we can also target $\mathbf{B}$ directly by maximizing the objective function $L(\mathbf{V})$. The downside of this approach is that $\mathbf{X}$ needs to be standardized (see toy example in Section 2.3) requiring the inversion of $\mathbf{\Sigma_X}$. Our choice of targeting the orthogonal complement avoids the inversion of $\mathbf{\Sigma_X}$, and the estimation algorithm in Section 2.6 can formally be applied to regressions with $p > n$ or $p \approx n$, where n denotes the sample size. Nevertheless, since the focus of this thesis is on a classic sufficient dimension reduction setting with $n > p$, we do not explore this further. Additionally, targeting the complement

has theoretical advantages. The proof of Theorem 12 reveals that the convergence rate is faster if $q = \dim(\mathbf{V})$ increases.

We conclude this Section in the next remark highlighting the difference of CVE to MAVE as both are based on a conditional variance.

Remark. *The difference of* CVE *to* MAVE *is that the objective function of* CVE *on the population level is given by* (2.2)*, whereas for* MAVE *it is given by* (1.40)*. Both are based on a conditional variance of the response, but* MAVE *conditions on the projection* $\mathbf{V}^T\mathbf{X}$ *and* CVE *conditions on* $\mathbf{X}$ *only ranging in a subspace* $\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$*. Moreover,* MAVE *estimates the forward model and reduction simultaneously using local linear smoothing and targets the reduction* $\mathbf{B}$ *in model* (1.18) *directly whereas* CVE *targets the orthogonal complement, i.e.* $\operatorname{span}\{\mathbf{B}\}^\perp$ *and circumvents the estimation of the forward model..*

2.2 Estimation of cve

So far we have defined the estimation equation on the population level in (2.20). To calculate the *Conditional Variance Estimator* from a sample of model (1.18) we replace the objective function, (2.2), in (2.18) by an estimate. This Section describes the estimation of the objective function (2.2) and the definition of the CVE estimator is given in Definition 15. The replacement of the unknown quantities g and $f_{\mathbf{X}}$ in (2.5), contained in the objective function (2.2), by standard nonparametric kernel estimates is unsuitable for the sufficient dimension reduction task since the goal is to avoid a nonparametric estimation over a high input dimension suffering from the curse of dimensionality. Therefore, we opted to use the kernel estimation approach, described below, considering the structure of the conditioning subspaces since this results in the effective dimension $p - q$ (which is substantially smaller than p, if k is small) over which the nonparametric smoothing takes place.

Assume $(Y_i, \mathbf{X}_i^T)_{i=1,\dots,n}^T$ is an independent identical distributed sample from model (1.18). For $\mathbf{V} \in \mathcal{S}(p,q)$ and $\mathbf{s}_0 \in \operatorname{supp}(f_{\mathbf{X}})$, we define

$$\begin{aligned} d_i(\mathbf{V}, \mathbf{s}_0) &= \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}}\mathbf{X}_i\|^2 = \|\mathbf{X}_i - \mathbf{s}_0\|^2 - \langle \mathbf{X}_i - \mathbf{s}_0, \mathbf{V}\mathbf{V}^T(\mathbf{X}_i - \mathbf{s}_0)\rangle \\ &= \|(\mathbf{I}_p - \mathbf{V}\mathbf{V}^T)(\mathbf{X}_i - \mathbf{s}_0)\|^2 = \|\mathbf{P}_{\mathbf{U}}(\mathbf{X}_i - \mathbf{s}_0)\|^2 \end{aligned} \tag{2.21}$$

where $\langle \cdot, \cdot \rangle$ is the usual inner product in $\mathbb{R}^p$, $\mathbf{P}_{\mathbf{V}} = \mathbf{V}\mathbf{V}^T$ and $\mathbf{P}_{\mathbf{U}} = \mathbf{I}_p - \mathbf{P}_{\mathbf{V}}$ using the orthogonal decomposition given by (1.3).

Let $h_n \in \mathbb{R}_+$ be a sequence of bandwidths and we call the set $\mathcal{S}_{\mathbf{s}_0,\mathbf{V}} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\operatorname{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$ a *slice* that depends on both the shifting point $\mathbf{s}_0$ and the matrix $\mathbf{V}$. h_n represent the squared width of a slice around the subspace $\mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$ and fulfills the following assumptions.

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

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

Remark. *(H.1) will guarantee that the bias goes to* 0 *and (H.2) that the variance does as well. For obtaining the consistency of the proposed estimator (H.2) will be strengthened to* $\log(n)/n h_n^{(p-q)/2} \to 0$.

Let K be a function satisfying the following assumptions.

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

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

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

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

Examples of functions that satisfy (K.1) and (K.2) include the Gaussian, $K(z) = c\exp(-z^2/2)$, the exponential, $K(z) = c\exp(-z)$, and the squared Epanechnikov kernel, $K(z) = c\max\{(1-z^2), 0\}^2$ (i.e. polynomial kernels), where c is a constant. The rectangular, $K(z) = cI(z \leq 1)$, does not fulfill the assumptions but will be mentioned for intuitive explanations. A list of further kernel functions is given in [Par61, Table 1].

2.2.1 The estimator of $L(\mathbf{V})$

Definition 12. *For $i = 1, \ldots, n$, we define*

$$w_i(\mathbf{V}, \mathbf{s}_0) = \frac{K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{d_j(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)} \tag{2.22}$$

Definition 13. *The sample based estimate of $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ is defined as*

$$\tilde{L}_n(\mathbf{V}, s_0) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0)(Y_i - \bar{y}_1(\mathbf{V}, \mathbf{s}_0))^2 = \bar{y}_2(\mathbf{V}, \mathbf{s}_0) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0)^2 \tag{2.23}$$

where $\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0) Y_i^l$, $l = 1, 2$.

Definition 14. *The estimate of the objective function $L(\mathbf{V})$ in (2.2) is defined as*

$$L_n(\mathbf{V}) = \frac{1}{n} \sum_{i=1}^n \tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \tag{2.24}$$

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

To obtain insight, see also Section 2.3, as to the choice of $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ in (2.23), let us consider the rectangular kernel, $K(z) = 1_{\{z \leq 1\}}$. In this case, $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ computes the empirical variance of the Y_i's corresponding to the $\mathbf{X}_i$'s that are no further than $\sqrt{h_n}$ away from the affine space $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, i.e., $d_i(\mathbf{V}, \mathbf{s}_0) = \|\mathbf{X}_i - \mathbf{P}_{\mathbf{s}_0 + \text{span}\{\mathbf{V}\}} \mathbf{X}_i\|^2 \leq h_n$. If a smooth kernel is used, such as the Gaussian in our simulation studies, then $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ is also smooth, which allows the computation of gradients required to solve the optimization problem.

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

In this section, we describe a slight adaptation of the estimation of the target function (2.2) that accounts for the fact that in the different slices, $\mathcal{S}_{\mathbf{s}_0,\mathbf{V}} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$, usually there are different amount of points.

In the estimation of $L(\mathbf{V})$ two different weighting schemes are used:

(a) *Within a slice.* The weights are defined in (2.22) and are used to calculate (2.23).

(b) *Between slices.* Equal weights $1/n$ are used to calculate (2.24).

The choice of weights can be potentially influential. Especially the between weighting scheme can further be refined by assigning more weight to slices with more points. This can be realized by altering (2.24) to

$$L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^{n} \tilde{w}(\mathbf{V}, \mathbf{X}_i)\tilde{L}_n(\mathbf{V}, \mathbf{X}_i), \quad \text{with} \tag{2.25}$$

$$\tilde{w}(\mathbf{V}, \mathbf{X}_i) = \frac{\sum_{j=1}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n) - 1}{\sum_{l,u=1}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n) - n} = \frac{\sum_{j=1,j\neq i}^{n} K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)}{\sum_{l,u=1,l\neq u}^{n} K(d_l(\mathbf{V}, \mathbf{X}_u)/h_n)} \tag{2.26}$$

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 $\tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$. Therefore this slice gets higher weight, if the number of $\mathbf{X}_j$ points in this slice, $\mathcal{S}_{\mathbf{X}_i,\mathbf{V}}$, is larger. That is, the more observations we use for estimating $L(\mathbf{V}, \mathbf{X}_i)$ the better its accuracy. The denominator in (2.26) guarantees the weights $\tilde{w}(\mathbf{V}, \mathbf{X}_i)$ sum up to one.

If (2.24) is replaced by (2.25) in Definition 15, the resulting estimator is called *weighted conditional variance estimator*.

2.3 Intuition of cve via an toy example

In this section we provide an intuitive explanation of the proposed method and demonstrate how the sample version (2.24) of the objective function (2.2) estimates the orthogonal complement of $\mathbf{B}$ in (1.18) via an example. We consider a bivariate normal predictor vector, $\mathbf{X} = (X_1, X_2)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_\mathbf{x})$. We generate the response from $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon = X_1 + \epsilon$, with $\epsilon \sim N(0, \eta^2)$ independent of $\mathbf{X}$ with $\eta = 0.1$. In this setting, $k = 1$, $\mathbf{B} = (1, 0)^T$, $g(z) = z \in \mathbb{R}$ in model (1.18), i.e. $\mathbf{B}$ is aligned with the first coordinate axis. Further, we set $\mathbf{\Sigma}_\mathbf{x} = \mathbf{I}_2$ for convenience.

First we draw a sample of size $n = 100$ and plot the $\mathbf{X}_i, i = 1, \ldots, n$ in Figure 2.1, where the color of the points are determined by their corresponding Y_i values, i.e. the low Y_i values are assigned blue and the higher the Y_i value the more red the points are. In the direction of $\mathbf{B}$, i.e. first axis left to right, the color has high variation, whereas in the direction $(0, 1)$, i.e. second axis up and down, the color has low variation only due to the error term ϵ.

For a given direction $\mathbf{V} \in \mathbb{R}^2$ we demonstrate the concept of (2.23), the left panel of Figure 2.1 uses $\mathbf{V} = \mathbf{B} = (1, 0)^T$ and the right panel $\mathbf{V} = (0, 1)^T \perp \mathbf{B}$ both with

shift point $\mathbf{s}_0 = (0,0)^T$ denoted as black cross. The subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, i.e. one dimensional line, is indicated via the black arrow and the black dotted lines represent the slice $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$. For each point $\mathbf{X}_i$, $d_i(\mathbf{V}, \mathbf{s}_0)$ in (2.21) is the squared distance to the subspace $\mathbf{s}_0+\text{span}\{\mathbf{V}\}$ and, if the rectangular kernel $K(z) = 1_{\{|z|\leq 1\}}$ is used, $K(d_i(\mathbf{V}, \mathbf{s}_0)/h_n)$ is 0 for points outside of the slice and 1 inside. Then $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ in (2.23) calculates the empirical variance of the Y_i values whose $\mathbf{X}_i$ values fall into the slice. In the left panel $\mathbf{V}$ is aligned with $\mathbf{B}$ which yields $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) = 1.034$, i.e. the color in the slice is very heterogeneous and has high variation, whereas in the left panel $\mathbf{V}$ is orthogonal to $\mathbf{B}$ resulting in $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) = 0.11$, i.e. the color in the slice is very homogeneous with low variation.

Since one slice uses only a fraction of the data and the variation in one slice only depends on the alignment of the direction $\mathbf{V}$ to $\mathbf{B}$ but not on the shifting point $\mathbf{s}_0$, as long as the slice is not too sparsely populated by samples, it is useful to average (2.23) over different shifting points $\mathbf{s}_0$ in order to use all data available. A convenient choice for shift points are the datapoints $\mathbf{X}_i$, therefore we average over all $\mathbf{X}_i$ in (2.24) to form the final estimate of the objective function.

The population quantity in (2.1) is given if the width of the slice, h_n, is infinitesimal small and $\mathbf{X}$ ranges only in the subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$, i.e. we have infinitely many $\mathbf{X}_i$ points that are directly on the line spanned by the black arrow. Then in the right panel (2.1) calculates $\mathbb{V}\text{ar}(\epsilon) = \eta^2$ and in the left panel $\mathbb{V}\text{ar}(Y \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}) + \mathbb{V}\text{ar}(\epsilon)$. This demonstrates Theorem 9.

Moreover, for this toy model we can calculate (2.2) explicitly via (2.3) and (2.5). We calculate for an arbitrary $\mathbf{\Sigma}_\mathbf{x}$ to demonstrate that the argmin of (2.2) is always orthogonal to $\mathbf{B}$ but the argmax can be influenced by the covariance matrix. With these specifications, (2.5) becomes

$$\mu_l(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}} (\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}r)^l f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(r)dr \tag{2.27}$$

Dropping the terms that do not contain $\mathbf{r}$ in (2.3) yields

$$\begin{aligned} f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(r) &\propto f_\mathbf{X}(\mathbf{s}_0 + \mathbf{V}r) \propto \exp\left(-\frac{1}{2}(\mathbf{s}_0 + r\mathbf{V})^T\mathbf{\Sigma}_\mathbf{x}^{-1}(\mathbf{s}_0 + r\mathbf{V})\right) \\ \propto \exp\left(-\frac{1}{2}\left(2r\mathbf{V}^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{s}_0 + r^2\mathbf{V}^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{V}\right)\right) &= \exp\left(-\frac{1}{2\sigma^2}\left(2r\sigma^2\mathbf{V}^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{s}_0 + r^2\right)\right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(r-\alpha)^2\right), \end{aligned} \tag{2.28}$$

where $\sigma^2 = 1/(\mathbf{V}^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{V})$, $\alpha = -\sigma^2\mathbf{V}^T\mathbf{\Sigma}_\mathbf{x}^{-1}\mathbf{s}_0$ and the symbol $\propto$ stands for proportional to. Letting $\psi(z)$ denote the density of a standard normal variable, (2.28) obtains

$$f_{\mathbf{X}|\mathbf{X}\in\mathbf{s}_0+\text{span}\{\mathbf{V}\}}(r) = \frac{1}{\sigma}\psi\left(\frac{r-\alpha}{\sigma}\right) \tag{2.29}$$

for $\mathbf{V}, \mathbf{s}_0 \in \mathbb{R}^{2\times 1}$. Inserting (2.29) in (2.27) yields

$$\int_{\mathbb{R}} (\mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}r)^l \frac{1}{\sigma}\psi\left(\frac{r-\alpha}{\sigma}\right)dr = \begin{cases} \mathbf{B}^T\mathbf{s}_0 + \mathbf{B}^T\mathbf{V}\alpha & l = 1 \\ (\mathbf{B}^T\mathbf{s}_0)^2 + 2(\mathbf{B}^T\mathbf{s}_0)(\mathbf{B}^T\mathbf{V})\alpha + (\mathbf{B}^T\mathbf{V})^2(\sigma^2+\alpha^2) & l = 2 \end{cases}$$

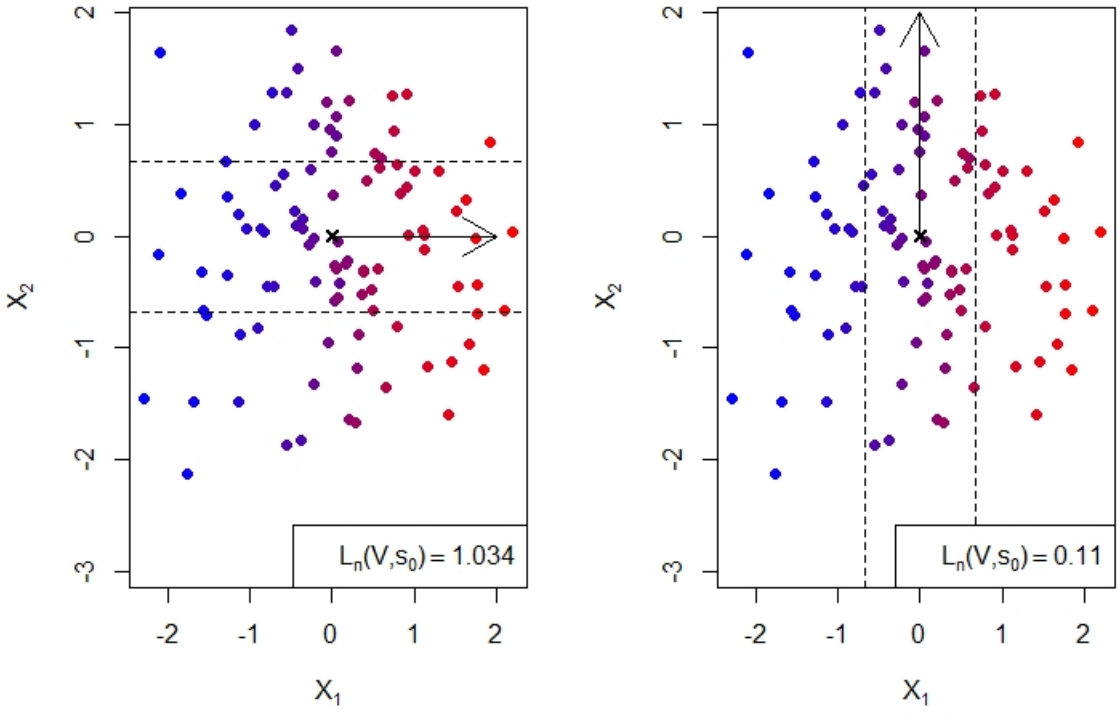


Figure 2.1: Plot of the $\mathbf{X}_i$ samples from toy model with $n = 100$, the color of the points are determined by their corresponding Y_i values, i.e. the low Y_i values are assigned blue and the higher the Y_i value the more red the points are. For the left panel $\mathbf{V} = \mathbf{B} = (1, 0)^T$, and for the right panel $\mathbf{V} = (0, 1)^T \perp \mathbf{B}$ both with shift point $\mathbf{s}_0 = (0, 0)^T$ denoted as black cross. The subspace $\mathbf{s}_0 + \text{span}\{\mathbf{V}\}$ is indicated via the black arrow and the black dotted lines represent the slice $\{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\mathbf{s}_0+\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h_n\}$.

Using (2.4) and (2.2), yields $\tilde{L}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)^2 + \eta^2 = (\mathbf{B}^T\mathbf{V})^2\sigma^2 + \eta^2$, so that

$$L(\mathbf{V}) = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right) = (\mathbf{B}^T\mathbf{V})^2\sigma^2 + \eta^2 = \frac{(\mathbf{B}^T\mathbf{V})^2}{\mathbf{V}^T\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathbf{V}} + \eta^2 \tag{2.30}$$

From (2.30) we can easily see that $L(\mathbf{V})$ attains its minimum at $\mathbf{V} \perp \mathbf{B}$. Also, if $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{I}_2$, the maximum of $L(\mathbf{V})$ is attained at $\mathbf{V} = \mathbf{B}$. To visualize the behavior of $\tilde{L}_n(\mathbf{V})$ as the sample size increases, we parametrize $\mathbf{V}$ by $\mathbf{V}(\theta) = (\cos(\theta), \sin(\theta))^T$, $\theta \in [0, \pi]$. Since $\mathbf{B} = (1, 0)^T$, the minimum of $L(\mathbf{V})$ is at $\mathbf{V}(\pi/2) = (0, 1)^T$, which is orthogonal to $\mathbf{B}$.

The true $L(\mathbf{V}(\theta))$ and its estimates $L_n(\mathbf{V}(\theta))$ are plotted for samples of different sizes n in Figure 2.2. $L_n(\mathbf{V}(\theta))$ approximates $L(\mathbf{V})$ fast and attains its minimum close to the same value as $L(\mathbf{V})$, even for $n = 10$, for this specific sample.

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

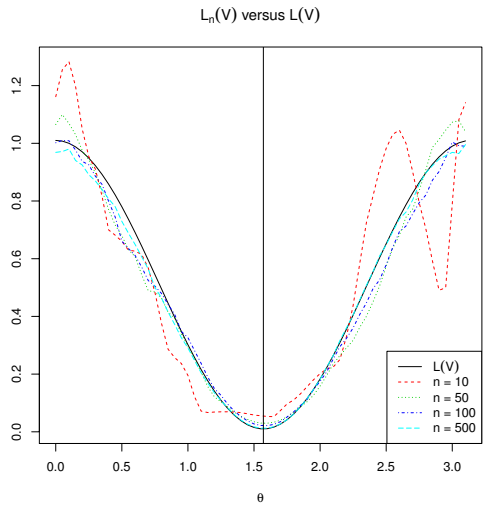


Figure 2.2: Solid black line is $L(\mathbf{V}(\theta)) = \cos(\theta)^2 + 0.1^2$, colored is $L_n(\mathbf{V}(\theta))$, $\theta \in [0, \pi]$, $n = 10, 50, 100, 500$. The vertical black line is at $\theta = \pi/2$

Alternatively, we can calculate the target function of the toy model $Y = \mathbf{B}^T\mathbf{X} + \epsilon$ with $\mathbf{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ a p-dimensional covariate vector independent to $\epsilon \sim N(0, \eta^2)$, and $\mathbf{B} \in \mathbb{R}^p$. Let $\mathbf{V} \in \mathcal{S}(p, q)$ and $\mathbf{U} \in \mathcal{S}(p, p-q)$ with $\mathbf{V} \perp \mathbf{U}$, then

$$\begin{aligned}\tilde{L}(\mathbf{V}, \mathbf{s}_0) &= \mathbb{V}\mathrm{ar}(Y \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) = \mathbb{V}\mathrm{ar}(\mathbf{B}^T\mathbf{X} \mid \mathbf{U}^T\mathbf{X} = \mathbf{U}^T\mathbf{s}_0) + \eta^2 \\ &= \mathbf{B}^T\mathbb{V}\mathrm{ar}(\mathbf{X} \mid \mathbf{U}^T\mathbf{X} = \mathbf{U}^T\mathbf{s}_0)\mathbf{B} + \eta^2 = \mathbf{B}^T\left(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{U}(\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Sigma}\right)\mathbf{B} + \eta^2\end{aligned}$$

where the last equality follows from the properties of the normal distribution, i.e.

$$\mathbf{X} \mid \mathbf{U}^T\mathbf{X} \sim N(\ldots, \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{U}(\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U})^{-1}\mathbf{U}^T\boldsymbol{\Sigma}).$$

Equation (3) of Appendix A in [CF09] states for any full rank matrix $\mathbf{M}$, we have

$$\mathbf{V}(\mathbf{V}^T\mathbf{M}\mathbf{V})^{-1}\mathbf{V}^T + \mathbf{M}^{-1}\mathbf{U}(\mathbf{U}^T\mathbf{M}^{-1}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{M}^{-1} = \mathbf{M}^{-1}$$

Using the equation above with $\mathbf{M} = \mathbf{\Sigma}^{-1}$,yields

$$\begin{aligned}\tilde{L}(\mathbf{V}, \mathbf{s}_0) &= \mathbf{B}^T\left(\mathbf{\Sigma} - \mathbf{\Sigma}\mathbf{U}(\mathbf{U}^T\mathbf{\Sigma}\mathbf{U})^{-1}\mathbf{U}^T\mathbf{\Sigma}\right)\mathbf{B} + \eta^2 \\ &= \mathbf{B}^T\mathbf{V}(\mathbf{V}^T\mathbf{\Sigma}^{-1}\mathbf{V})^{-1}\mathbf{V}^T\mathbf{B} + \eta^2\end{aligned}$$

which agrees with (2.30) for $p = 2$. Further, since in this model $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ does not depend on $\mathbf{s}_0$ (i.e. the variance of a conditional normal distribution is constant with respect to the conditioning argument) we have $L(\mathbf{V}) = \mathbb{E}\left(\tilde{L}(\mathbf{V}, \mathbf{X})\right) = \tilde{L}(\mathbf{V}, \mathbf{s}_0)$.

Connection of cve to lad: Using formula Equation (3) of Appendix A in [CF09], which is a building stone in the proofs of Prop 1 and 2 of [CF09] used to formulate the LAD estimator (see Section 1.10), is a tool to calculate the target function of CVE on the population level for the toy model with linear link function and normal distributed $\mathbf{X}$.

2.4 Bandwidth selection

The performance of conditional variance estimation depends crucially on the choice of the bandwidth sequence h_n that controls the bias-variance trade-off if the mean squared error is used as measure for accuracy, in the sense that the smaller h_n is, the lower the bias and the higher the variance and vice versa. Furthermore, the choice of h_n depends on p, q, the sample size n, and the distribution of $\mathbf{X}$. We assume throughout the bandwidth satisfies assumptions (H.1) and (H.2). We will use Lemma 11 to derive a possible bandwidth rule.

Lemma 11. *Let $\mathbf{M}$ be a $p \times p$ positive definite matrix. Then,*

$$\frac{tr(\mathbf{M})}{p} = \operatorname{argmin}_{s>0} \|\mathbf{M} - s\mathbf{I}_p\| \tag{2.31}$$

Proof. Let $\mathbf{U}$ be the $p \times p$ matrix whose columns are the eigenvectors of $\mathbf{M}$ corresponding to its eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p > 0$. Then, $\mathbf{M} = \mathbf{U}\mathrm{diag}(\lambda_1, ..., \lambda_p)\mathbf{U}^T$, which implies $\|\mathbf{M} - s\mathbf{I}_p\|_2^2 = \|\mathrm{diag}(\lambda_1, ..., \lambda_p) - s\mathbf{I}_p\|^2 = \sum_{l=1}^p (\lambda_l - s)^2$. Taking the derivative with respect to s, setting it to 0 and solving for s obtains (2.31), since $\sum_{l=1}^p \lambda_l = \mathrm{tr}(\mathbf{M})$. □

If the predictors are multivariate normal, their joint density is approximated by $N(\mu_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$ by Lemma 11, with $\sigma^2 = \mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})/p$. This results in no bandwidth dependence on $\mathbf{V}$ and leads to a rule for bandwidth selection, as follows.

Under $\mathbf{X} \sim N_p(\mu_{\mathbf{X}}, \sigma^2\mathbf{I}_p)$, $\widetilde{\mathbf{X}}_i = \mathbf{X}_i - \mathbf{X}_j \sim N_p(0, 2\sigma^2\mathbf{I}_p)$ for $i \neq j$, where we suppress the dependence on j for notational convenience. Since all data are used as shifting points, $d_i(\mathbf{V}, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|^2 - (\mathbf{X}_i - \mathbf{X}_j)^T\mathbf{V}\mathbf{V}^T(\mathbf{X}_i - \mathbf{X}_j) = \|\widetilde{\mathbf{X}}_i\|^2 - \widetilde{\mathbf{X}}_i^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}}_i$. Let

$$\begin{aligned}\mathrm{nObs} &= \mathbb{E}\left(\#\{i \in \{1, ..., n\} : \widetilde{\mathbf{X}}_i \in \mathrm{span}_h\{\mathbf{V}\}\}\right) \\ &= 1 + (n-1)\mathbb{P}(d_1(\mathbf{V}, \mathbf{X}_2) \leq h) = 1 + (n-1)\mathbb{P}(\|\widetilde{\mathbf{X}}\|^2 - \widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} \leq h)\end{aligned} \tag{2.32}$$

where $\text{span}_h\{\mathbf{V}\} = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{P}_{\text{span}\{\mathbf{V}\}}\mathbf{x}\|^2 \leq h\}$ and $\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{X}^*$, with $\mathbf{X}^*$ an independent copy of $\mathbf{X}$. nObs is the expected number of points in a slice. Given a user specified value for nObs, h is the solution to (2.32).

Let $\mathbf{x} \in \mathbb{R}^p$. For any $\mathbf{V} \in \mathcal{S}(p,q)$ in (1.1), there exists an orthonormal basis $\mathbf{U} \in \mathbb{R}^{p\times(p-q)}$ of $\text{span}\{\mathbf{V}\}^\perp$ such that $\mathbf{x} = \mathbf{V}\mathbf{r}_1 + \mathbf{U}\mathbf{r}_2$, by (1.3). Then, $\widetilde{\mathbf{X}} = \mathbf{V}\mathbf{R}_1 + \mathbf{U}\mathbf{R}_2$, with $\mathbf{R}_1 = \mathbf{V}^T\widetilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_q), \mathbf{R}_2 = \mathbf{U}^T\widetilde{\mathbf{X}} \sim N(0, 2\sigma^2\mathbf{I}_{p-q})$, and $\widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} = \|\mathbf{R}_1\|^2$ and $\|\widetilde{\mathbf{X}}\|^2 = \|\mathbf{R}_1\|^2 + \|\mathbf{R}_2\|^2$. Therefore,

$$\mathbb{P}\left(\|\widetilde{\mathbf{X}}\|^2 - \widetilde{\mathbf{X}}^T\mathbf{V}\mathbf{V}^T\widetilde{\mathbf{X}} \leq h\right) = \mathbb{P}(\|\mathbf{R}_2\|^2 \leq h) = \chi_{p-q}\left(\frac{h}{2\sigma^2}\right), \tag{2.33}$$

where χ_{p-q} is the cumulative distribution function of a chi-squared random variable with $p-q$ degrees of freedom. Plugging (2.33) in (2.32) obtains

$$\text{nObs} = 1 + (n-1)\chi_{p-q}\left(\frac{h}{2\sigma^2}\right). \tag{2.34}$$

Solving (2.34) for h and Lemma 11 yield

$$h_n(\text{nObs}) = \chi^{-1}_{p-q}\left(\frac{\text{nObs}-1}{n-1}\right)\frac{2\text{tr}(\widehat{\Sigma}_\mathbf{x})}{p}, \tag{2.35}$$

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

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

Alternatively, a plug-in bandwidth based on rule-of-thumb rules of the form $csn^{-1/(4+k)}$, where s is an estimate of scale and c a number close to 1, such as Silverman's ($c = 1.06$, s =standard deviation) or Scott's ($c = 1$, s =standard deviation), used in nonparametric density estimation [see [Sil86]], is

$$h_n = 1.2^2\frac{2\text{tr}(\widehat{\Sigma}_\mathbf{x})}{p}\left(n^{-1/(4+p-q)}\right)^2. \tag{2.36}$$

The term $2\text{tr}(\widehat{\Sigma}_\mathbf{X})/p$ can be interpreted as the variance of $\mathbf{X}_i - \mathbf{X}_j$ and $p-q$ is the true dimension k. We use 1.2 as c based on empirical evidence from simulations. Since both (2.35) and (2.36) yield satisfactory results, we opted against cross validation for bandwidth selection because of the computational burden involved, and used the bandwidth in (2.36) in simulations and data analyses.

2.5 Consistency of cve

In this chapter some asymptotic results and the consistency of CVE are presented. Theorem 12 states the conditions under which $L_n(\mathbf{V})$ in (2.24) converges uniformly in probability to its population counterpart in (2.2). This result will lead to the consistency of CVE in Theorem 13.

Theorem 12. *Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1),* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$ *, and* $a_n/h_n^{(p-q)/2} = O(1)$*,*

$$\sup_{\mathbf{V}\in\mathcal{S}(p,q)} |L_n(\mathbf{V}) - L(\mathbf{V})| \to 0 \quad \textit{in probability as } n \to \infty \tag{2.37}$$

The proof of Theorems 12 and 13 are deferred to chapter 4 due to historical reasons. Both proofs represent a main part of the theoretical contributions of this thesis but since the inception of CVE we already worked out the generalisation presented in chapter 4. Theorems 12 is a special cases of Theorem 25 and follows immediately from it if the ensemble only containing the identity function is used.

Definition 15. *The sample based* ***Conditional Variance Estimator*** $\widehat{B}_{p-q}$ *is any basis of* $\operatorname{span}\{\widehat{\mathbf{V}}_q\}^\perp$ *where* $\widehat{\mathbf{V}}_q = \operatorname{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L_n(\mathbf{V})$.

Next we establish the consistency of the conditional variance estimator. The uniform convergence in probability of the sample objective function in (2.24) is a sufficient condition for obtaining the consistency of $\widehat{\mathbf{V}}_q = \operatorname{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L_n(\mathbf{V})$, as uniform convergence in probability of a random function implies convergence in probability of the minimizer of $L_n(\mathbf{V})$ to the minimizer of the limit function. The main theoretical result follows in Theorem 13 which establishes the consistency of CVE.

Theorem 13. *Under (A.1), (A.2), (A.3), (A.4), (K.1), (K.2), (H.1),* $a_n^2 = \log(n)/nh_n^{(p-q)/2} = o(1)$*,* *and* $a_n/h_n^{(p-q)/2} = O(1)$*,* $\operatorname{span}\{\widehat{\mathbf{B}}_k\}$ *is a consistent estimator for* $\operatorname{span}\{\mathbf{B}\}$ *in model (1.18); i.e.,*

$$\|\mathbf{P}_{\widehat{\mathbf{B}}_k} - \mathbf{P}_{\mathbf{B}}\| \to 0 \quad \textit{in probability as } n \to \infty.$$

The proof of Theorem 13 follows immediately from Theorem 26.

2.6 Optimization Algorithm

A Stiefel manifold optimization algorithm is used to obtain the solution of the sample version of the optimization problem (2.18). To calculate $\widehat{\mathbf{V}}_q$ in (15), a curvilinear search is carried out [WY13, WY10, Tag11], which is similar to gradient descent. First an arbitrary starting value $\mathbf{V}^{(0)}$ is selected by drawing a $p\times q$ matrix from the invariant measure; i.e., the distribution that corresponds to the uniform, on $\mathcal{S}(p,q)$, see [Chi03]. The Q-component of the QR decomposition of a $p \times q$ matrix with independent standard normal entries follows the invariant measure [Chi94]. The step-size $\tau > 0$, the step size reduction factor $\gamma \in (0,1)$, and tolerance $\text{tol} > 0$ are fixed at the outset.

Result: $\mathbf{V}^{(\text{end})}$
Initialize: $\mathbf{V}^{(0)}$, $\tau = 1$, $\text{tol} = 10^{-3}$, $\gamma = 0.5$ $\text{error} = \text{tol} + 1$, $\text{maxit} = 50$, $\text{count} = 0$;
while *error* $>$ *tol and count* $\leq$ *maxit* **do**

- $\mathbf{G} = \nabla_{\mathbf{V}} L_n(\mathbf{V}^{(j)}) \in \mathbb{R}^{p \times q}$, $\mathbf{W} = \mathbf{G}\mathbf{V}^T - \mathbf{V}\mathbf{G}^T$
- $\mathbf{V}^{(j+1)} = (\mathbf{I}_p + \tau \mathbf{W})^{-1}(\mathbf{I}_p - \tau \mathbf{W})\mathbf{V}^{(j)}$
- $\text{error} = \|\mathbf{V}^{(j)}\mathbf{V}^{(j)\mathbf{T}} - \mathbf{V}^{(j+1)}\mathbf{V}^{(j+1)^T}\|/\sqrt{2q}$

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

Algorithm 1: Curvilinear search

Under mild regularity conditions on the objective function, [WY13] showed that the sequence generated by the algorithm converges to a stationary point if the Armijo-Wolfe conditions [NW06] are used for determining the stepsize τ.

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

The algorithm is repeated for m arbitrary $\mathbf{V}^{(0)}$ starting values drawn from the invariant measure on $\mathcal{S}(p, q)$. Among those, the value at which L_n in (2.24) is minimal is selected as $\widehat{\mathbf{V}}_q$.

The algorithm requires the computation of the gradient of $L_n(\mathbf{V})$ in (2.24) or (2.25). We compute the gradient of the objective function for the Gaussian kernel in Theorems 14 and 15. The Gaussian kernel is the default kernel we use in the implementation of the estimation algorithm in the `R` code that accompanies this manuscript.

Theorem 14. *Let* $K(z) = \exp(-z^2/2)$ *be the Gaussian kernel. Then, the gradient of* $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ *in* (2.23) *is given by*

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

and the gradient of $L_n(\mathbf{V})$ *in* (2.24) *is*

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

with $w_i = w(\mathbf{V}, \mathbf{X}_i)$ *in* (2.22).

Proof of Theorem 14. The Gaussian kernel K satisfies $\partial_z K(z) = -zK(z)$. From (2.22) and (2.23) we have $\tilde{L}_n = \bar{y}_2 - \bar{y}_1^2$ where $\bar{y}_l = \sum_i w_i Y_i^l$, $l = 1, 2$. 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, $\nabla_{\mathbf{V}} K_i = (-1/h_n^2) K_i d_i \nabla_{\mathbf{V}} d_i$ and $\nabla_{\mathbf{V}} w_i = -\left(K_i d_i \nabla_{\mathbf{V}} d_i (\sum_j K_j) - K_i \sum_j K_j d_j \nabla_{\mathbf{V}} d_j\right) / (h_n \sum_j K_j)^2$. Next,

$$\begin{aligned}
\nabla_{\mathbf{V}} \bar{y}_l &= -\frac{1}{h_n^2} \sum_i Y_i^l \frac{\left(K_i d_i \nabla_{\mathbf{V}} d_i - K_i (\sum_j K_j d_j \nabla_{\mathbf{V}} d_j)\right)}{(\sum_j K_j)^2} \\
&= -\frac{1}{h_n^2} \sum_i Y_i^l w_i \left(d_i \nabla_{\mathbf{V}} d_i - \sum_j w_j d_j \nabla_{\mathbf{V}} d_j \right) \\
&= -\frac{1}{h_n^2} \left(\sum_i Y_i^l w_i d_i \nabla_{\mathbf{V}} d_i - \sum_j Y_j^l w_j \sum_i w_i d_i \nabla_{\mathbf{V}} d_i \right) = -\frac{1}{h_n^2} \sum_i (Y_i^l - \bar{y}_l) w_i d_i \nabla_{\mathbf{V}} d_i
\end{aligned} \tag{2.38}$$

Then, $\nabla_{\mathbf{V}} \tilde{L}_n = \nabla_{\mathbf{V}} \bar{y}_2 - 2\bar{y}_1 \nabla_{\mathbf{V}} \bar{y}_1$, and inserting $\nabla_{\mathbf{V}} \bar{y}_l$ from (2.38) yields $\nabla_{\mathbf{V}} \tilde{L}_n = (-1/h_n^2) \sum_i (Y_i^2 - \bar{y}_2 - 2\bar{y}_1(Y_i - \bar{y}_1)) w_i d_i \nabla_{\mathbf{V}} d_i = (1/h_n^2)(\sum_i \left(\tilde{L}_n - (Y_i - \bar{y}_1)^2\right) w_i d_i \nabla_{\mathbf{V}} d_i)$, since $Y_i^2 - \bar{y}_2 - 2\bar{y}_1(Y_i - \bar{y}_1) = (Y_i - \bar{y}_1)^2 - \tilde{L}_n$. □

The weighted version of conditional variance estimation in Section 2.2.2 is expected to increase the accuracy of the estimator for unevenly spaced data. When (2.25) and the gradient in (2.39) are used in the optimisation algorithm, we refer to the estimator as *weighted conditional variance estimation*. If (2.25) and the gradient $\sum_{i=1}^n \tilde{w}(\mathbf{V}, \mathbf{X}_i) \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$ is used; i.e., the first summand in (2.39) is dropped, we refer to it as *partially weighted conditional variance estimation*. For both, we replace G in algorithm 1 with the corresponding gradient derived in Theorem 15.

Theorem 15. *Let* $K(z) = \exp(-z^2/2)$ *be the Gaussian kernel. Then, the gradient of* $L_n^{(w)}(\mathbf{V})$ *in* (2.25) *is given by*

$$\nabla_{\mathbf{V}} L_n^{(w)}(\mathbf{V}) = \sum_{i=1}^{n} \left(\nabla_{\mathbf{V}} \tilde{w}(\mathbf{V}, \mathbf{X}_i) \tilde{L}_n(\mathbf{V}, \mathbf{X}_i) + \tilde{w}(\mathbf{V}, \mathbf{X}_i) \nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i) \right), \tag{2.39}$$

where $\nabla_{\mathbf{V}} \tilde{L}_n(\mathbf{V}, \mathbf{X}_i)$ *is given in Theorem 14. Furthermore,*

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

with $\tilde{w}_i = \tilde{w}(\mathbf{V}, \mathbf{X}_i)$ *in* (2.26), $K_{j,i} = K(d_j(\mathbf{V}, \mathbf{X}_i)/h_n)$, *and* $d_{j,i} = d_j(\mathbf{V}, \mathbf{X}_i)$ *given in* (2.21).

The proof of Theorem 15 is analogue to the proof of Theorem 14 and will not be presented.

CVE has computational complexity of $O(m\texttt{maxit}pqn^2)$ as can be seen by Theorem 14 since for each entry of the $p \times q$ dimensional gradient a double sum with both indices ranging over $1, \ldots, n$ has to be evaluated, and for the m starting values maximal `maxit` iterations are used. The formulas of the gradient are straightforward to implement but there is no closed form solution available due to the nonlinear nature, i.e. the argument $\mathbf{V}$ is inside of the nonlinear kernel. The CVE method is efficiently implemented in the `R` package `CVarE` [KF21] available at `cran`. Nevertheless, for usual sample sizes the efficient MAVE implementation in the `R` package [WY19] is usually a bit faster but the difference in the speed advantage is not dramatic. Moreover, the computational complexity of CVE can be controlled by the number of arbitrary starting values for the curvilinear search used in the optimization procedure.

2.7 Simulations

2.7.1 Simulation Study: Demonstrating the consistency

Moreover we explore the consistency of the *conditional variance estimator* (CVE) through a simulation study. The model is given by:

$$Y = (\mathbf{b}_1^T \mathbf{X})(\mathbf{b}_2^T \mathbf{X}) + 0.5\epsilon \tag{2.40}$$

where $p = 12$, $k = 2$, $\mathbf{X} \sim N(0, I_{12})$, $\epsilon \sim N(0, 1)$ independent of $\mathbf{X}$,

$\mathbf{b}_1 = (1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0)^T/6^{1/2} \in \mathbb{R}^{12}$, and $\mathbf{b}_2 = (1, -1, 1, -1, 1, -1, 0, 0, 0, 0, 0, 0)^T/6^{1/2} \in \mathbb{R}^{12}$.

The *conditional variance estimator* is compared with the forward model based sufficient dimension reduction methods, meanMAVE (`meanMAVE`) [XTLZ02], central subspace MAVE (`csMAVE`) [WX08] and `pHd` [Li92, CL02], and the inverse regression based method, SIR [Li91]. Central subspace MAVE (`csMAVE`) assumes model (1.13) $Y = g(\mathbf{B}^T\mathbf{X}, \epsilon)$, which is a much more general model than (1.18). The reference methods `meanMAVE` and `csMAVE` are implemented in the R package MAVE, `pHd` and SIR are implemented in the dr package. The dimension $k = 2$ and $q = 10$ are assumed to be known throughout. The Conditional Variance estimator is used with four different choices for the bandwidth h_n, they will be denoted CVE1, CVE2, CVE3 and CVE4. The first three use the bandwidth choice proposed in (2.35), i.e. $h_n(\text{nObs}) = 2\chi^{-1}_{p-q}((\text{nObs}-1)/(n-1))\text{tr}(\Sigma_{\mathbf{X}})/p$ with $\text{nObs} = n^{4/5}$ for CVE1, $\text{nObs} = n^{2/3}$ for CVE2, and $\text{nObs} = n^{1/2}$ for CVE3. The bandwidth for CVE4 is given by the plug-in rule (2.36), i.e. $h_n = 2(\text{tr}(\Sigma_{\mathbf{X}})/p)(1.2/n^{1/(4+p-q)})^2$ where $\text{tr}(\Sigma_{\mathbf{X}})$ is estimated as the trace of the maximum likelihood estimate of the covariance-matrix $\Sigma_{\mathbf{X}}$.

The simulation is performed by:

For a given sample size n, 100 i.i.d samples $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}^j$ for $j = 1, .., 100$ are drawn from (2.40), then for each method $\mathbf{B}$ is estimated from sample j and then $\text{err}_{j,n} = \|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|/(2k)^{1/2}$ is calculated. This is done for sample sizes $(25, 50, 100, 200, 300, 400, 600, 800)$

and Figure 2.3 displays the distribution of $err_{j,n}$ for increasing n for the different methods. Figure 2.3 indicates that except from SIR all methods are consistent in model (2.40). In

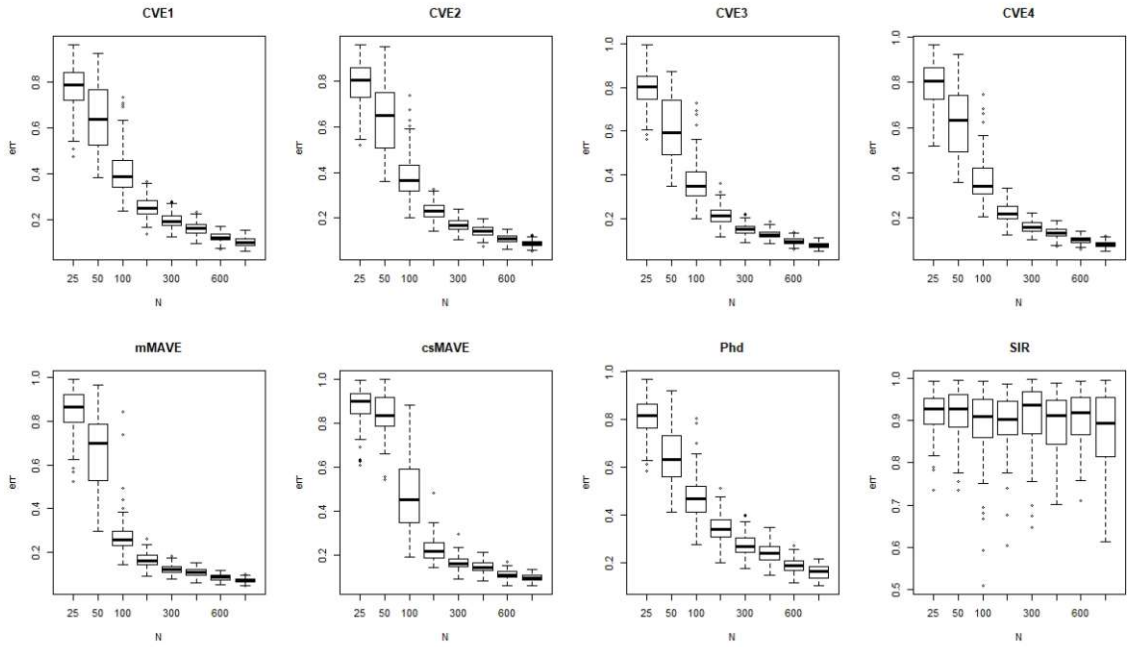


Figure 2.3: Estimation error distribution plotted over $n = (25, 50, 100, 200, 300, 400, 600, 800)$ for the different methods

Figure 2.4 the $\log(\bar{\text{err}}_n)$ values are plotted against $\log(n)$ where $\bar{\text{err}}_n = \sum_{j=1}^{100} \text{err}_{j,n}/100$ is an estimate for $\mathbb{E}(\|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|)/\sqrt{2k}$. If the mean error $\mathbb{E}(\|\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|) \approx Cn^\alpha$ decays with a power law, then one expects to see lines in Figure 2.4. Figure 2.4 indicates that for large n the rate of convergence is similar for all methods and that `meanMAVE` has the lowest mean estimation error for $n > 50$.

To get a clearer picture we repeat the simulations with model 2.40 from above for CVE4 (denoted as CVE), meanMAVE, and pHd with increased sample sizes in Figure 2.5,

i.e. $n \in \{25, 50, 100, 200, 300, 400, 600, 800, 1000, 1500, 2000\}$. Figure 2.5 displays $\log(\bar{\text{err}}_n)$ against $\log(n)$ and for $n = 2000$ ($\log(2000) \approx 7.6$) CVE achieves a slightly lower mean estimation error than meanMAVE. Moreover, Figure 2.5 could indicate that CVE has a slightly faster convergence rate than meanMAVE for certain models.

2.7.2 Simulations to evaluate estimation accuarcy

We compare the estimation accuracy of conditional variance estimation with the forward model based sufficient dimension reduction methods, mean outer product gradient estimation (`meanOPG`), mean minimum average variance estimation (`meanMAVE`) [WY19], refined outer product gradient (`rOPG`), refined minimum average variance estimation (`rmave`) [XTLZ02, Li18], and principal Hessian directions (`pHd`) [Li92, CL02], and the inverse regression based methods, sliced inverse regression (`SIR`) [Li91] and sliced average variance

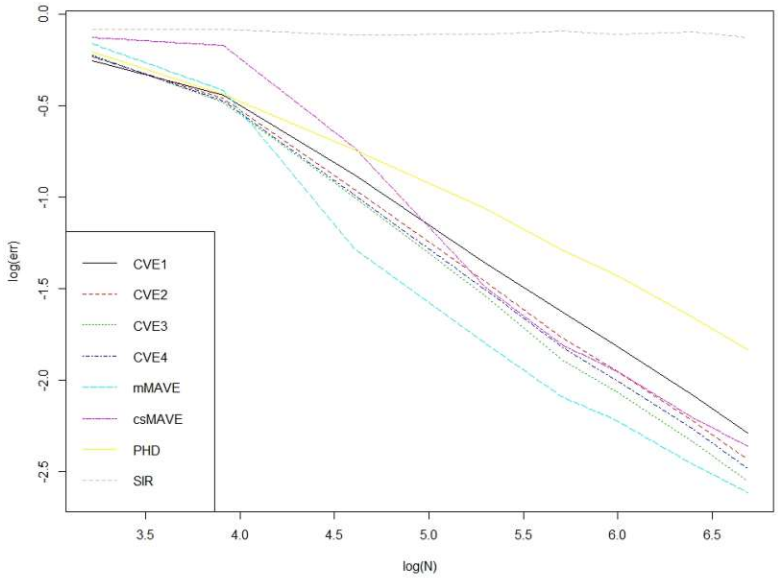


Figure 2.4: $\log(\bar{\text{err}}_n)$ for different method plotted against $\log(n)$ for $n \in (25, 50, 100, 200, 300, 400, 600, 800)$

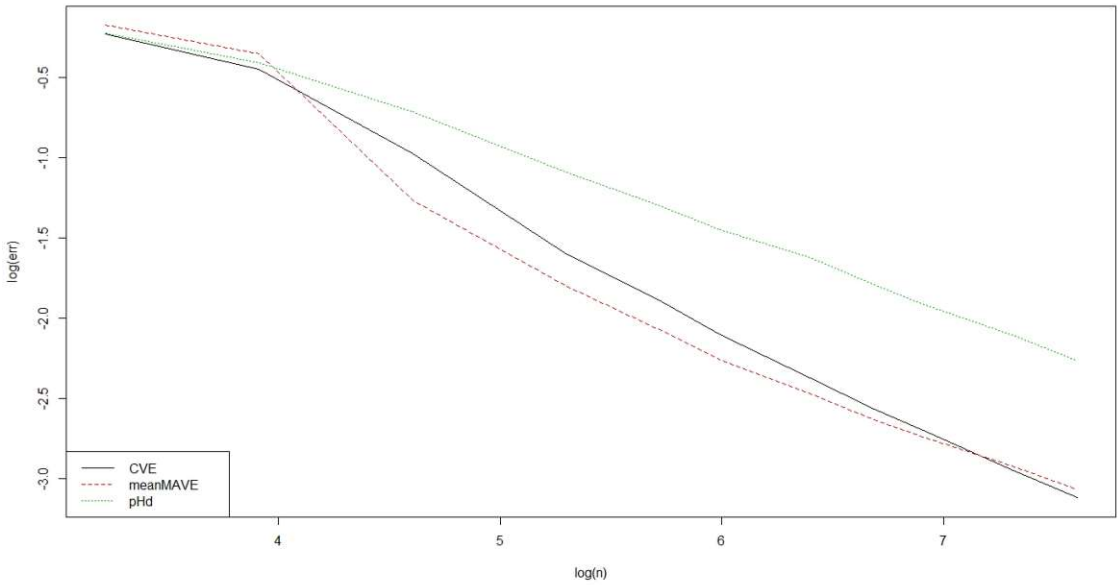


Figure 2.5: $\log(\bar{\text{err}}_n)$ for CVE4, meanMAVE, and pHd against $\log(n)$ for $n \in (25, 50, 100, 200, 300, 400, 600, 800, 1000, 1500, 2000)$

estimation (SAVE) [CW91]. The dimension k is assumed to be known throughout.

We report results for conditional variance estimation using the "plug-in" bandwidth in (2.36) and three different conditional variance estimation versions, CVE, wCVE, and rCVE. CVE is obtained by using $m = 10$ arbitrary starting values in the optimization algorithm and optimizing (2.24) as described in Section 2.6. rCVE, or *refined weighted CVE*, is obtained by setting the starting value $\mathbf{V}^{(0)}$ at the optimizer of CVE, and using (2.25) in the optimization algorithm in Section 2.6 with the partially weighted gradient as described in Section 2.2.2. wCVE, or *weighted CVE*, is obtained by optimizing (2.25) with partially weighted gradient as described in Sections 2.2.2 and 2.6. Methods rOPG and rmave refer to the original refined outer product gradient and refined minimum average variance estimation algorithms published in [XTLZ02]. They are implemented using the R code in [Li18] with number of iterations nit = 25, since the algorithm is seen to converge by 25. The dr package is used for the SIR, SAVE and pHd calculations, and the MAVE package for mean outer product gradient estimation (meanOPG) and mean minimum average variance estimation (meanMAVE). The source code for conditional variance estimation can be downloaded from https://git.art-ist.cc/daniel/CVE and is also available in the R package CVarE.

Table 2.1 lists the seven models (M1-M7) we consider. Throughout, we set $p = 20$, $\mathbf{b}_1 = (1,1,1,1,1,1,0,...,0)^T/\sqrt{6}$, $\mathbf{b}_2 = (1,-1,1,-1,1,-1,0,...,0)^T/\sqrt{6} \in \mathbb{R}^p$ for M1-M5. For M6, $\mathbf{b}_1 = \mathbf{e}_1, \mathbf{b}_2 = \mathbf{e}_2$ and $\mathbf{b}_3 = \mathbf{e}_p$, and for M7 $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are the same as in M6 and $\mathbf{b}_4 = \mathbf{e}_3$, where $\mathbf{e}_j$ denotes the p-vector with jth element equal to 1 and all others are 0. The error term ϵ is independent of $\mathbf{X}$ for all models. In M2, M3, M4, M5 and M6, $\epsilon \sim N(0,1)$. For M1 and M7, ϵ has a generalized normal distribution $GN(a,b,c)$ with densitiy $f_\epsilon(z) = c/(2b\Gamma(1/c))\exp((|z-a|/b)^c)$, see [Nad05] with location 0 and shape-parameter 0.5 for M1, and shape-parameter 1 for M7 (Laplace distribution). For both the scale-parameter is chosen such that $\mathbb{V}\mathrm{ar}(\epsilon) = 0.25$.

Table 2.1: Models

Name	Model	$\mathbf{X}$ distribution	ϵ distribution	k	n
M1	$Y = \cos(\mathbf{b}_1^T\mathbf{X}) + \epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$	$GN(0, \sqrt{1/2}, 0.5)$	1	100
M2	$Y = \cos(\mathbf{b}_1^T\mathbf{X}) + 0.5\epsilon$	$\mathbf{X} \sim \lambda Z\mathbf{1}_p + N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	1	100
M3	$Y = 2\log(\lvert\mathbf{b}_1^T\mathbf{X}\rvert + 2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	1	100
M4	$Y = (\mathbf{b}_1^T\mathbf{X})/(0.5 + (1.5 + \mathbf{b}_2^T\mathbf{X})^2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	$N(0,1)$	2	200
M5	$Y = \cos(\pi\mathbf{b}_1^T\mathbf{X})(\mathbf{b}_2^T\mathbf{X} + 1)^2 + 0.5\epsilon$	$\mathbf{X} \sim U([0,1]^p)$	$N(0,1)$	2	200
M6	$Y = (\mathbf{b}_1^T\mathbf{X})^2 + (\mathbf{b}_2^T\mathbf{X})^2 + (\mathbf{b}_3^T\mathbf{X})^2 + 0.5\epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	3	200
M7	$Y = (\mathbf{b}_1^T\mathbf{X})(\mathbf{b}_2^T\mathbf{X})^2 + (\mathbf{b}_3^T\mathbf{X})(\mathbf{b}_4^T\mathbf{X}) + \epsilon$	$\mathbf{X} \sim t_3(\mathbf{I}_p)$	$GN(0, \sqrt{1/\Gamma(6)}, 1)$	4	400

The variance-covariance structure of $\mathbf{X}$ in models M1 and M4 satisfies $\mathbf{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $i,j = 1,\ldots,p$. In M5, $\mathbf{X}$ is uniform with independent entries on the p-dimensional hyper-cube. In M7, $\mathbf{X}$ is multivariate t-distributed with 3 degrees of freedom. The link functions of M4 and M7 are studied in [XTLZ02], but we use $p = 20$ instead of 10 and a non identity covariance structure for M4 and the t-distribution instead of normal for M7. In M2, $Z \sim 2\text{Bernoulli}(p_{\text{mix}}) - 1 \in \{-1,1\}$, where $\mathbf{1}_q = (1,1,...,1)^T \in \mathbb{R}^q$, mixing probability $p_{\text{mix}} \in [0,1]$ and dispersion parameter $\lambda > 0$. For $0 < p_{\text{mix}} < 1$, $\mathbf{X}$ has a mixture normal distribution, where p_{mix} is the relative mode height and λ is a measure of mode distance.

We set $q = p - k$ and generate $r = 100$ replications of models M1 - M7. We estimate

$\mathbf{B}$ using the ten sufficient dimension reduction methods. The accuracy of the estimates is assessed using $err = \|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|/\sqrt{2k}$, which lies in the interval $[0, 1]$. The factor $\sqrt{2k}$ normalizes the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement. This is a consequence of $\|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|^2 = \|\mathbf{P}_{\mathbf{B}}\|^2 - 2\text{tr}(\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\widehat{\mathbf{B}}}) + \|\mathbf{P}_{\widehat{\mathbf{B}}}\|^2 \leq \|\mathbf{P}_{\mathbf{B}}\|^2 + \|\mathbf{P}_{\widehat{\mathbf{B}}}\|^2 = \text{tr}(\mathbf{P}_{\mathbf{B}}^2) + \text{tr}(\mathbf{P}_{\widehat{\mathbf{B}}}^2) = 2\text{tr}(\mathbf{I}_k) = 2k$ since $\text{tr}(\mathbf{P}_{\mathbf{B}}\mathbf{P}_{\widehat{\mathbf{B}}}) \geq 0$ for projections.

In Table 2.2 the mean and standard deviation of *err* for M1 - M7 are reported. In particular, for M2, $p_{mix} = 0.3$ and $\lambda = 1$. The smallest error values are boldfaced. In models M1, M2 and M3, the conditional variance estimator is the best performer, with its refined version as close second. In M4, M5 and M6, any of the four versions of MAVE performs better than the CVE. For model M7 the results of `rOPG` and `rmave` are not reported because the code frequently produces an error message that a matrix is not invertible. Among the rest, the weighted version of CVE, wCVE, attains the minimum error.

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

In Figure 2.6, box-plots for all combinations of $p_{\text{mix}} \in \{0.3, 0.4, 0.5\}$ and $\lambda \in \{0, 0.5, 1, 1.5\}$ are presented. The reference methods are restricted to `meanOPG` and `meanMAVE`, since the others are not competitive. Conditional variance estimation performs better than all competing methods and is the only method with consistently smaller errors when the two modes are further apart ($\lambda \geq 1$) regardless of the mixing probability p_{mix}. The performance of both `meanOPG` and `meanMAVE` worsens as one moves from left to right row-wise. The mixing probability, p_{mix}, has no noticeable effect on the performance of any method; i.e., the plots are very similar column-wise. In sum, `meanMAVE`'s performance deteriorates as the bimodality of the predictor distribution becomes more distinct. In contrast, conditional variance estimation is unaffected and appears to have an advantage over `meanMAVE` when the predictors have mixture distributions, the link function is even about the midpoint of the two modes, and $\mathbf{B}$ is not orthogonal to the line connecting the two modes. Conditional variance estimation is the only method that estimates the mean subspace reliably in model M2 ($err \approx 0.4$ to 0.5), whereas `meanMAVE` misses it completely ($err \approx 1$). These results indicate that conditional variance estimation is often approximately on par, and can perform much better than `meanMAVE` depending on the predictor distribution and the link function.

Furthermore we estimate the dimension k via cross-validation, following the approach in [XTLZ02] , with

$$\hat{k} = \text{argmin}_{l=1,\ldots,p}\, CV(l) = \text{argmin}_{l=1,\ldots,p} \frac{\sum_i (Y_i - \hat{g}^{-i}(\widehat{\mathbf{B}}_l^T \mathbf{X}_i))^2}{n}, \tag{2.41}$$

where $\hat{g}^{-i}(\cdot)$ is computed from the data $(Y_j, \widehat{\mathbf{B}}_l^T \mathbf{X}_j)_{j=1,\ldots,n; j\neq i}$ using multivariate adaptive regression splines [Fri91] in the `R`-package `mda`, and $\widehat{\mathbf{B}}_l = \widehat{\mathbf{V}}_{p-l}^{\perp}$ is any basis of the orthogonal complement of $\widehat{\mathbf{V}}_{p-l} = \text{argmin}_{\mathbf{V}\in\mathcal{S}(p,p-l)} L_n(\mathbf{V})$. For a given l, we calculate $\widehat{\mathbf{B}}_l$ from the whole data set and predict Y_i by $\hat{Y}_{i,l} = \hat{g}^{-i}(\widehat{\mathbf{B}}_l^T \mathbf{X}_i)$. For $l = p$, $\widehat{\mathbf{B}}_p = \mathbf{I}_p$. The results for the seven models are reported in Table 2.3. The CVE based dimension estimation is the most accurate in models M1, M2, M3, and M6 and differs slightly from that of MAVE in M7. MAVE performs better in M4 and M5, completely misses the true dimension in

Table 2.2: Mean and standard deviation of estimation errors

Model		CVE	wCVE	rCVE	meanOPG	rOPG	meanMAVE	rmave	pHd	SIR	SAVE
$M1$											
	mean	**0.3827**	0.4414	0.4051	0.6220	0.9876	0.5099	0.9840	0.8278	0.9875	0.9788
	sd	0.1269	0.1595	0.1329	0.1879	0.0223	0.1800	0.0295	0.1206	0.0243	0.0334
$M2$											
	mean	**0.4572**	0.4992	0.4658	0.8987	0.9332	0.8905	0.9242	0.9000	0.9783	0.9781
	sd	0.1038	0.1524	0.0989	0.0908	0.0683	0.0983	0.0897	0.0735	0.0278	0.0318
$M3$											
	mean	**0.6282**	0.7509	0.6371	0.7847	0.9644	0.7576	0.9674	0.6964	0.9647	0.9519
	sd	0.2354	0.2262	0.2181	0.2201	0.0667	0.2435	0.0609	0.1626	0.0587	0.0650
$M4$											
	mean	0.5663	0.5897	0.5554	0.4071	0.4026	0.4361	**0.3905**	0.7772	0.5824	0.9727
	sd	0.1239	0.1246	0.1298	0.0814	0.0609	0.0997	0.0584	0.0662	0.0951	0.0202
$M5$											
	mean	0.4429	0.5604	0.4779	0.4058	**0.3737**	0.3929	0.3750	0.7329	0.6374	0.9730
	sd	0.0891	0.1233	0.0976	0.1022	0.0680	0.0894	0.0871	0.0832	0.0968	0.0186
$M6$											
	mean	0.3828	0.3027	0.3230	0.1827	0.4632	**0.1656**	0.4863	0.4978	0.9129	0.8236
	sd	0.1006	0.0748	0.1098	0.0289	0.1717	0.0252	0.1676	0.0601	0.0420	0.0518
$M7$											
	mean	0.6856	**0.5050**	0.5651	0.5694	NA	0.5482	NA	0.8536	0.8133	0.8699
	sd	0.0588	0.0862	0.0879	0.1122	NA	0.1271	NA	0.0354	0.0341	0.0342

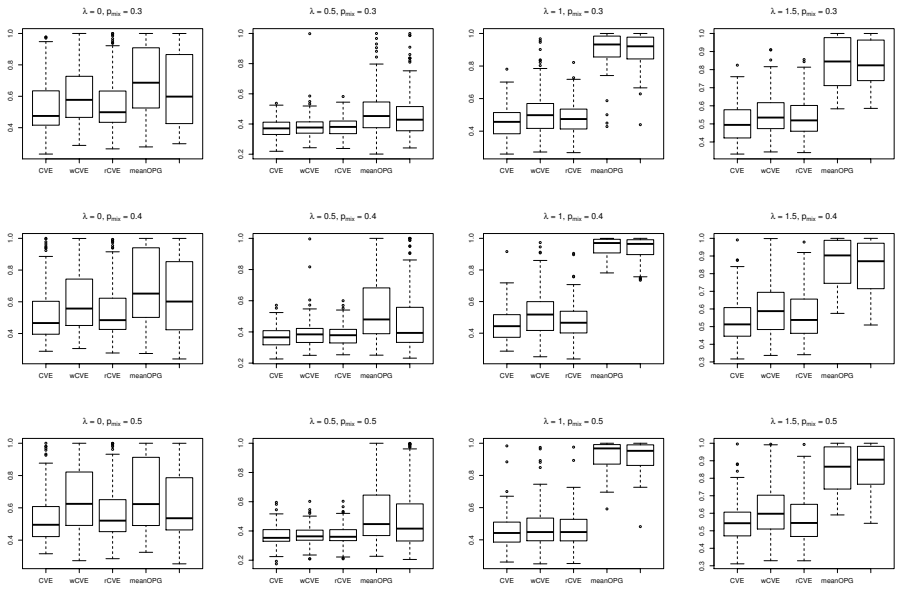


Figure 2.6: M2, $p = 20, n = 100$

M2 and misses it most of the time in M3. Thus, the dimension estimation performance of CVE and MAVE agrees with the estimation accuracy of the true subspace in Table 2.2, CVE estimates the dimension more accurately even in model M6, where it exhibits worse subspace estimation performance, and overall appears to be more accurate.

Table 2.3: Number of times dimension k is correctly estimated in 100 replications

	M1	M2	M3	M4	M5	M6	M7
CVE	83	41	88	62	46	74	19
MAVE	67	0	14	76	60	57	21

We carried out many simulation experiments for an array of combinations of link functions, sufficient reduction matrices $\mathbf{B}$ and their ranks, as well as predictor and error distributions. All reported and unreported results indicate that the difference in performance of the two methods, CVE and mean MAVE, can be attributed to both the form of the link function and the marginal predictor distribution. We observed that when the link function had a bounded first order derivative, CVE often outperformed meanMAVE across predictor distributions. In the opposite case, MAVE performed mostly better. Also, when the predictors have a bimodal distribution with well separated modes and the link function is even, regardless of whether its derivative is bounded, CVE outperforms meanMAVE. In the other settings for the generated data, both methods were roughly on par.

2.8 Data Analysis

Three data sets are analyzed: the *Hitters* data in the `R` package `ISLR`, which was also analyzed by [XTLZ02], the *Boston Housing* data in the `R` package `mlbench`, and the *Concrete* data from the `MAVE` package. The reference method is `meanMAVE` from the `MAVE` package in `R` and the CVE is calculated using $m = 50$ and maxit $= 10$ in the optimization algorithm 1 in Section 2.6. The estimation of the dimension is based on (2.41) in Section 2.7.

Following [XTLZ02], we remove 7 outliers from the *Hitters* data set leading to a sample size of 256. The response is $Y = \log(\text{salary})$ and the 16 continuous predictors are the game statistics of players in the Major League Baseball league in the seasons 1986 and 1987. Further information can be found in https://www.rdocumentation.org/packages/ISLR/versions/1.2/topics/Hitters.

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

The *Concrete* data set contains 1030 instances on 9 continuous variables The response is concrete compressive strength. Concrete strength is very important in civil engineering and is a highly nonlinear function of age and ingredients. The description of the variables

can be found in https://www.rdocumentation.org/packages/MAVE/versions/1.3.10/topics/Concrete.

For all three data sets we standardize both the predictors and the response by subtracting the mean and rescaling column-wise so that each variable has unit variance. The data sets are analyzed using 10 fold cross-validation to calculate an unbiased estimate of the prediction error [Sto74] for our method , CVE, and its main competitor `meanMAVE` using the `MAVE` package. The dimension for each method is estimated with (2.41) on the trainings set and we then fit a forward regression model on the training set replacing the original with the reduced predictors using multivariate adaptive regression splines [Fri91] using the `R` package `mda` and calculate the prediction error on the test set for both methods. The dimension estimates of CVE and `MAVE` mostly disagree.

The mean and standard deviation of the 10-fold cross-validation prediction errors are reported in Table 2.4. Since the response is standardized, the values in Table 2.4 are in the range of 0 to 1, with smaller values indicating better predictive performance. CVE performs slightly worse than mean `MAVE` in the *Hitters* data set, slightly better in the *Boston Housing* and better in the *Concrete* data set analysis.

Table 2.4: Mean and standard deviation (in parenthesis) of standardized out of sample prediction errors for the three data sets

Method	Hitters	Housing	Concrete
CVE	0.216	0.260	0.361
	(0.101)	(0.331)	(0.206)
MAVE	0.203	0.299	0.417
	(0.083)	(0.382)	(0.348)

2.8.1 Hitters Data Analysis as in [XTLZ02]

Additionally, we reconstruct the analysis of the *Hitters* data in [XTLZ02], which does not account for the out-of-sample prediction error as in Section 2.8 but uses the whole sample for estimation of $\mathbf{B}$ and its rank. Only the dimension k is estimated with leave-one-out cross validation.

Table 2.5 reports the average cross validation mean squared error $CV(k)$ in (2.41) using the whole data set over $k = 1, \ldots, 5$. Both conditional variance estimation and mean minimum average variance estimation estimate the dimension to be 2.

Table 2.5: Mean cross-validation error

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

We plot the response against the estimated directions in Figure 2.7. Both exhibit the

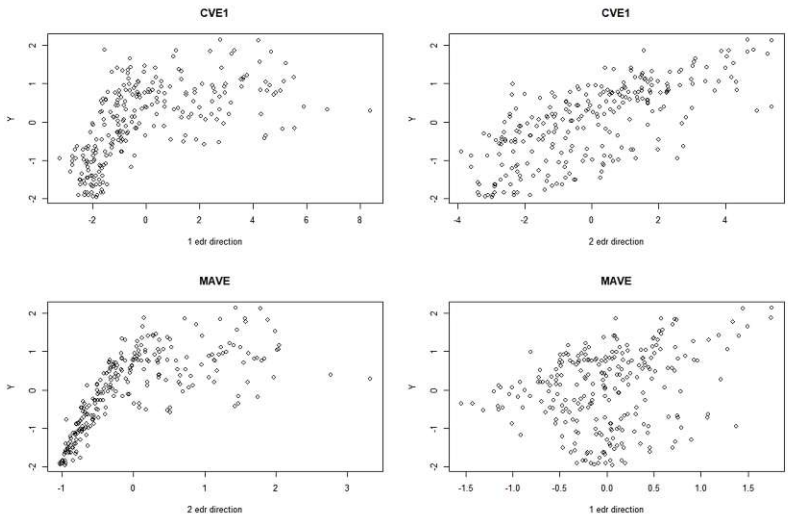


Figure 2.7: Y against $\widehat{\mathbf{b}}_1^T\mathbf{X}$ and $\widehat{\mathbf{b}}_2^T\mathbf{X}$

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

Based on the scatterplots in Figure 2.7, we fit the same models for both. For conditional variance estimation, the fitted regression is

$$\hat{Y} = 0.39578 + 0.33724(\widehat{\mathbf{b}}_1^T\mathbf{X}) - 0.08066(\widehat{\mathbf{b}}_1^T\mathbf{X})^2 + 0.29126(\widehat{\mathbf{b}}_2^T\mathbf{X}) \tag{2.42}$$

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

$$\hat{Y} = 0.39051 + 1.32529(\widehat{\mathbf{b}}_1^T\mathbf{X}) - 0.55328(\widehat{\mathbf{b}}_1^T\mathbf{X})^2 + 0.49546(\widehat{\mathbf{b}}_2^T\mathbf{X}) \tag{2.43}$$

with $R^2 = 0.7859$. Both models (2.42) and (2.43) have about the same fit as measured by R^2. The in sample performance of the two methods is practically the same for the `Hitters` data.

2.9 Discussion

In this chapter the novel conditional variance estimator (CVE) for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ is introduced. We present its geometrical and theoretical foundation, show its consistency and propose an estimation algorithm with assured convergence. CVE requires the forward model (1.18), $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$, holds and weak assumptions on the response and the covariates.

Minimum average variance estimation (MAVE) [XTLZ02] is the state of the art SDR method for the mean subspace and is the natural competitor. It estimates the sufficient dimension reduction targeting both the reduction and the link function g in (1.18). CVE targets only the reduction and does not require estimation of the link function, which may explain why it has an advantage over MAVE in some regression settings.

In Section 2.7 the performance of CVE is demonstrated via Simulations. They show that CVE is roughly on par to MAVE for most models and can yield substantial improvements in others. For example in the bimodal M2, CVE exhibits similar performance across different link functions (cos, exp, etc) for fixed λ, whereas the performance of MAVE is very uneven for M2 in Section 2.7. CVE is more accurate than MAVE when the link function is even and the predictor distribution is bimodal throughout our simulation studies. Moreover, CVE does not require the inversion of the predictor covariance matrix and can be applied to regressions with $p \approx n$ or $p > n$. The performance in such setting has not been explored and is a future line of research together with establishing more asymptotic properties like the rate of convergence or asymptotic distributions.

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

3 Neural Net SDR for the mean subspace

Most forward regression SDR methods, e.g. MAVE and CVE, are usable in relatively small p and n regression problems. When both p and n increase substantially, their computation can spread over days or weeks, thus rendering them infeasible in practice. Nowadays, many data applications easily exceed these thresholds.

In this Chapter 3 the novel NN−SDR estimator for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\}$ in (1.18) is introduced. The NN − SDR estimator combines forward regression SDR with neural networks in order to remove the limitation of small p and n. We propose a two stage NN − SDR estimator that carries out simultaneous sufficient dimension reduction and neural network learning.

Neural nets have become the "go-to" data analysis method in nearly all scientific fields due to their successful application, especially, for large sample sizes and input dimension in regression problems [Spe91, AKG19], image recognition [PHPP18, LB98, Fas02], speech recognition [AL91, KJB08] and many more applications [GBC16].

First we fit an arbitrary neural net to the data, and in the second stage we refine the estimate with a specific architecture using a bottleneck. The premise of the two stage NN − SDR estimator is conceptually similar to MAVE with the difference that we use neural nets as universal function approximators compared to nonparametric local linear smoothing methods. The advantage of this approach is that it retains the accuracy of state of the art SDR methods while it can be easily deployed to large scale datasets frequently encountered in applications. It also obtains predictions at nearly no additional computational cost compared to fully non-parametric methods used in MAVE and CVE. Further, the extension of the proposed NN − SDR estimator to online learning, where new data are dynamically added, is straightforward.

In Section 3.1 we present neural nets and the notation used throughout. In Section 3.2 we propose the novel two stage estimator and in Section 3.3 describe the algorithm. Then in Sections 3.4, 3.5 we draw the analogy to existing SDR methods and demonstrate its performance in Sections 3.6, 3.7, and 3.8 via simulations and data examples.

3.1 The Multi Layer Perceptron (MLP)

In this section we briefly review the concept of a *Multi Layer Perceptron* (`MLP` [Gur97, MP43, GBC16, JWHT14]) and introduce the corresponding notation.

An `MLP` is the concatenation of layers. Each layer consists of simple functions $f^{(l)}(x) = \phi(\mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)})$, where $\mathbf{W}^{(l)}$ is a matrix of *weights*, $\mathbf{b}^{(l)}$ is the *bias* vector of layer l, and together they form an affine transformation, on which the *activation* function $\phi(\cdot)$ is applied component-wise. The formal definition is provided next.

Definition 16. *A* Multi Layer Perceptron *(MLP) with N layers from $\mathbb{R}^p \to \mathbb{R}$ is a function with the following structure*

$$f_{MLP_N}(\mathbf{x}; \boldsymbol{\Theta}) = f^{(N)} \circ f^{(N-1)} \circ \ldots \circ f^{(1)}(\mathbf{x}) \tag{3.1}$$

where $\boldsymbol{\Theta} = (\mathbf{W}_1, \mathbf{b}_1, \ldots, \mathbf{W}_N, \mathbf{b}_N)$ and the l-th layer is given by

$$f^{(l)}(\mathbf{x}; \mathbf{W}^{(l)}, \mathbf{b}^{(l)}) = \phi^{(l)}(\mathbf{W}^{(l)}\mathbf{x} + \mathbf{b}^{(l)})$$

with weights $\mathbf{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$, bias $\mathbf{b}^{(l)} \in \mathbb{R}^{n_l}$, and a non-constant, continuous activation function $\phi^{(l)} : \mathbb{R} \to \mathbb{R}$ that is applied component-wise.

The notation $\boldsymbol{\Theta} = (\mathbf{W}_1, \mathbf{b}_1, \ldots, \mathbf{W}_N, \mathbf{b}_N)$ means that all parameters of an `MLP` are collected in vectorised form into the vector $\boldsymbol{\Theta} = (\text{vec}(\mathbf{W}_1), \mathbf{b}_1, \ldots, \text{vec}(\mathbf{W}_N), \mathbf{b}_N) \in \mathbb{R}^{\sum_{j=1}^{N} n_{l-1}n_l + n_l}$, where the operation $\text{vec} : \mathbb{R}^{n_{l-1} \times n_l} \to \mathbb{R}^{n_{l-1}n_l}$ stacks the columns of a matrix one after another.

The first layer that receives the input $\mathbf{x}$ is called the *input layer*, and the last layer is the *output layer*. All other layers are called *hidden layers*. A widely used activation function is the so called `ReLU` (Rectified Linear Unit) given by

$$\phi_{\texttt{ReLU}}(x) = \max(0, x).$$

The `ReLu` activation function will be used throughout this paper. Other popular choices include sigmoid functions like the tangens-hyperbolicus.

Figure 3.1 depicts a 3 layer `MLP`, $f_{\text{MLP}_3}(\mathbf{x}; \boldsymbol{\Theta})$, with input dimension 4; i.e., $\mathbf{x} = (x_1, \ldots, x_4)^T \in \mathbb{R}^4$. The first layer $f^{(1)}$ has output dimension 6, or 6 so called *neurons*, $\mathbf{W}_1 \in \mathbb{R}^{6\times4}, \mathbf{b}_1 \in \mathbb{R}^6$. The second layer, $f^{(2)}$, has 4 neurons with $\mathbf{W}_2 \in \mathbb{R}^{4\times6}, \mathbf{b}_2 \in \mathbb{R}^4$, and the output layer, $f^{(3)}$, has 1 neuron with $\mathbf{W}_3 \in \mathbb{R}^{1\times4}, \mathbf{b}_2 \in \mathbb{R}$. The arrows represent the weights of the layer. At each node (neuron), the bias is added before the activation function $\phi^{(l)}$ is applied.

The universal approximator theorem [Hor91, Thm 3] established that *Multi Layer Perceptrons* (`MLP`s) are universal approximators of functions. Theorem 16, which asserts that any continuously differentiable function can be approximated arbitrarily close on compact sets by an `MLP`, reproduces it.

Theorem 16. *Let $\texttt{MLP}_\infty$ be the set of all one layer `MLP`'s with arbitrarily many neurons in the first layer and the activation function ϕ is non-constant and bounded, then $\texttt{MLP}_\infty$ is uniformly m dense in $C^m(\mathbb{R}^p)$ on compact sets, where $C^m(\mathbb{R}^p)$ is the space of all m-times differentiable functions on $\mathbb{R}^p$.*

An application of Theorem 16 with $m = 1$, yields that for $g(\mathbf{B}^T\mathbf{x}) \in C^1(\mathbb{R}^p)$ of model (1.18), for every arbitrary compact set $K \subset \mathbb{R}^p$ and for all $\nu > 0$, there exists a one layer `MLP` $f_{\text{MLP}_1}(\cdot; \boldsymbol{\Theta})$ such that

$$\sup_{\mathbf{x}\in K} \left(|g(\mathbf{B}^T\mathbf{x}) - f_{\text{MLP}_1}(\mathbf{x}; \boldsymbol{\Theta})| + \|\nabla_{\mathbf{x}} g(\mathbf{B}^T\mathbf{x}) - \nabla_{\mathbf{x}} f_{\text{MLP}_1}(\mathbf{x}; \boldsymbol{\Theta})\| \right) \leq \nu$$

Therefore, the conditional expectation $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = g(\mathbf{B}^T\mathbf{x})$ and its gradients can be approximated arbitrarily close on compact sets by a one layer `MLP`. This serves as the motivation for the proposed estimation procedure in Section 3.2.

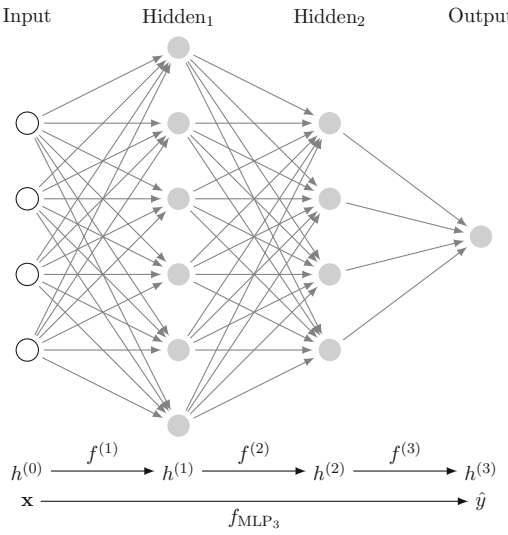


Figure 3.1: Example Architecture of a 3 Layer MLP.

3.2 nn − sdr Estimator

Theorems 7 and 6 present two ways of identifying $\mathbf{B}$ in model (1.18) at the population level. They serve as the motivation for the proposed NN − SDR estimator.

For $\mathbf{V} \in \mathcal{S}(p,k)$, recall $T(\mathbf{V})$ in (1.40) is the target function at the population level for MAVE and identifies $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ as shown in Theorem 6.

Next we define the three different neural nets we use in the proposed estimators.

Definition 17. *Using the notation for MLP's in Section 3.1, we define*

$$g_{\texttt{NN\_OPG}}(\mathbf{x};\boldsymbol{\Theta}_1) = f_{MLP_1}(\mathbf{x};\boldsymbol{\Theta}_1) : \mathbb{R}^p \to \mathbb{R} \tag{3.2}$$

$$g_{\texttt{NN\_wrap}}(\mathbf{x};\boldsymbol{\Theta}_2) = f_{MLP_1}(\mathbf{x};\boldsymbol{\Theta}_2) : \mathbb{R}^k \to \mathbb{R} \tag{3.3}$$

$$g_{\text{NN}}(\mathbf{x};(\mathbf{V},\boldsymbol{\Theta}_2)) = g_{\texttt{NN\_wrap}}(\mathbf{V}^T\mathbf{x};\boldsymbol{\Theta}_2) : \mathbb{R}^p \to \mathbb{R} \tag{3.4}$$

where $\mathbf{V} \in \mathcal{S}(p,k)$ is given in (1.1).

Our estimation method is run in two stages. The first uses $g_{\texttt{NN\_OPG}}(\mathbf{x};\boldsymbol{\Theta}_1)$ in (3.2) as an estimator for $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ in Theorem 7 . The second, or, *refinement* stage estimates g in model (1.18) with $g_{\texttt{NN\_wrap}}(\mathbf{x};\boldsymbol{\Theta}_2)$ in (3.3) and $g(\mathbf{B}^T\mathbf{x})$ in model (1.18) with $g_{\text{NN}}(\mathbf{x};(\mathbf{V},\boldsymbol{\Theta}_2))$ in (3.4). Both (3.2) and (3.4) are used to estimate $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x})$ but the latter uses the specific structure of model (1.18) for refinement.

The MLP in (3.4) is the same as in (3.3) save for an additional input layer with the identity as the activation function; i.e., $\phi^{(1)}(x) = x$. Further, the first layer forms a bottleneck, as depicted in Figure 3.2, since $\mathbf{V}^T\mathbf{x} \in \mathbb{R}^k$ with $k \ll p$. (3.2) serves as an estimate for $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = g(\mathbf{B}^T\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$, (3.3) for $g : \mathbb{R}^k \to \mathbb{R}$, and (3.4) is a refined estimate of $g(\mathbf{B}^T\mathbf{x})$ in model (1.18). The bottleneck of the `MLP` in (3.4) is conceptually similar to autoencoders [see, e.g., [KW19, Kra91]] with the important difference that the latter are

analogous to nonlinear principal components and unsupervised; that is, independent of the response.

Figure 3.2 illustrates the `MLP` in (3.4) with input dimension $p = 4$, $\mathbf{x} = (x_1, \ldots, x_4)^T \in \mathbb{R}^4$. The first layer represents the 2-dimensional linear reduction $\mathbf{V} \in \mathcal{S}(4, 2)$ and the rest of the network coincides with (3.3).

For the proposed estimator, we can use any `MLP` that has more neurons than p in the first hidden layer in (3.2) and (3.3). For the sake of simplicity we opted for a 1 layer `MLP` with 512 neurons as default, since this gave satisfactory results in simulations. Further, the performance in simulations was robust against different architectures if sufficient regularisation is applied via *dropout* in the training of the `MLP` [see [SHK+14]].

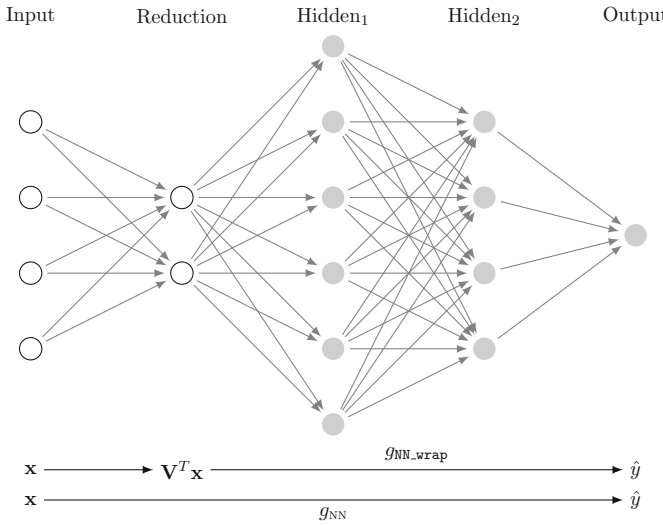


Figure 3.2: Illustration of g_{NN} in (3.4)

3.2.1 Initial Estimator

We assume $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$ is a random sample from the joint distribution of Y and $\mathbf{X}$ given by model (1.18). Let

$$T_{\texttt{NN\_OPG}}(\boldsymbol{\Theta}_1) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i, g_{\texttt{NN\_OPG}}(\mathbf{X}_i; \boldsymbol{\Theta}_1)\right) \tag{3.5}$$

be the objective function for the initial estimator, where $\mathcal{L} : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is a loss function. The training of the initial MLP in (3.2) is carried out by minimizing the objective function in (3.5),

$$\widehat{\boldsymbol{\Theta}}_1 = \text{argmin}_{\boldsymbol{\Theta}_1} \, T_{\texttt{NN\_OPG}}(\boldsymbol{\Theta}_1) \tag{3.6}$$

The resulting $g_{\texttt{NN\_OPG}}(\mathbf{x}, \widehat{\boldsymbol{\Theta}}_1)$ is an estimate of $\mathbb{E}(Y \mid \mathbf{X} = \mathbf{x}) = g(\mathbf{B}^T\mathbf{x})$ in model (1.18) if the squared error loss,

$$\mathcal{L}(x, y) = (x - y)^2 \,, \tag{3.7}$$

is used.

We set $\mathbf{b}_i = \nabla_{\mathbf{x}} g_{\texttt{NN\_OPG}}(\mathbf{X}_i, \widehat{\boldsymbol{\Theta}}_1) \in \mathbb{R}^p$ where $\widehat{\boldsymbol{\Theta}}_1$ is defined in (3.6), which is an estimate for $\nabla g(\mathbf{B}^T\mathbf{X}_i) \in \mathbb{R}^p$. We let

$$\widehat{\boldsymbol{\Sigma}}_{\texttt{NN\_OPG}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{b}_j \mathbf{b}_j{}^T \in \mathbb{R}^{p\times p} \tag{3.8}$$

that is an estimator for $\boldsymbol{\Sigma}_\nabla$ in Theorem 7. The `NN_OPG` estimator is defined as

$$\widehat{\mathbf{B}}_{\texttt{NN\_OPG}} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \in \mathcal{S}(p,k) \tag{3.9}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the first k eigenvectors of (3.8).

By Theorem 7, under model (1.18), $\text{span}\{\boldsymbol{\Sigma}_\nabla\} = \text{span}\{\mathbf{B}\} = \mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. If we assume that (3.8) is a consistent estimator for $\boldsymbol{\Sigma}_\nabla$ in Theorem 7, then the `NN_OPG` estimator in (3.9) is consistent for $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ in model (1.18). $\widehat{\mathbf{B}}_{\texttt{NN\_OPG}}$ in (3.9) is used as an initial starting value for the optimization in (3.11) in order to obtain the refined estimator $\widehat{\mathbf{B}}_{\text{NN}}$.

The loss function $\mathcal{L}$ is determined by model (1.18) and the conditional distribution of $Y \mid \mathbf{X}$. If the response Y and predictors $\mathbf{X}$ are continuous and the error term in (1.18) has a conditional Gaussian distribution, then the squared error loss function corresponds to the likelihood function. If Y is Bernoulli or multinomial distributed, then the cross entropy loss function can be used, and if Y is Poisson distributed then the deviance is the natural choice for the loss function. In general, the loss function is the relevant part of the likelihood in the conditional distribution of $Y \mid \mathbf{X}$ and agrees with the loss function in generalized linear models for conditional distributions in the exponential family.

3.2.2 Refinement Estimator

The second stage is the refinement of the initial estimator in (3.9). The `NN_OPG` estimator is obtained via the gradient of the trained MLP in (3.2). The training of the function $g_{\texttt{NN\_OPG}} : \mathbb{R}^p \to \mathbb{R}$ in (3.6) suffers from the curse of dimensionality if the input dimension is large. In this case, the accuracy of the estimation of (3.8) is adversely affected as learning a nonlinear function and its gradient with a high dimensional input space is difficult. The refinement procedure explicitly incorporates the defining assumption of model (1.18) that a lower dimension projection of the input, $\mathbf{B}^T\mathbf{X}$, can replace the original input $\mathbf{X}$. This is realised via the function g_{NN} in (3.4).

Definition 18. *The target function for the refinement estimator is given by*

$$T_{\text{NN}}(\mathbf{V}, \boldsymbol{\Theta}_2) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(Y_i, g_{\text{NN}}(\mathbf{X}_i; (\mathbf{V}; \boldsymbol{\Theta}_2))\right) \tag{3.10}$$

where $\mathcal{L} : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ *is a loss function, and* $\mathbf{V} \in \mathcal{S}(p,k)$. *Further, we set*

$$\left(\widehat{\mathbf{B}}_{\text{NN}}, \widehat{\boldsymbol{\Theta}}_2\right) = \text{argmin}_{\mathbf{V}\in\mathcal{S}(p,k),\boldsymbol{\Theta}_2}\, T_{\text{NN}}(\mathbf{V}, \boldsymbol{\Theta}_2) \tag{3.11}$$

and the NN − SDR *refinement estimator is given by* $\widehat{\mathbf{B}}_{\text{NN}}$.

The simultaneous optimization with respect to $\mathbf{V}$ and $\boldsymbol{\Theta}_2$ in (3.11) corresponds to simultaneous estimation of the sufficient reduction $\mathbf{B}$ and the link function g in model (1.18). The partially trained function $T_{\text{NN}}(\cdot, \widehat{\boldsymbol{\Theta}}_2)$ is an estimate for (1.39) if the squared error loss function (3.7) is used.

Under squared error loss, if (3.3) is a consistent estimator for g in (1.18), then $T_{\text{NN}}(\cdot, \widehat{\boldsymbol{\Theta}}_2)$ is consistent for $T(\mathbf{V})$ in (1.39). By Theorem 6, $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\} = \text{span}\{\text{argmin}_{\mathbf{V}\in\mathcal{S}(p,k)} T(\mathbf{V})\}$ and the expectation would be that, subject to regularity conditions, the NN refinement estimator is consistent.

Nevertheless, proving the consistency of (3.9) and the refinement estimator $\widehat{\mathbf{B}}_{\text{NN}}$ in (3.11) requires neural nets consistently estimate any function g from a sample of model (1.18). To the best of the authors' knowledge there is no such result available in the literature.

The optimization in (3.11) is solved via stochastic gradient descent training [see Section 3.3 or any other first order training algorithm for neural nets]. These algorithms require a starting value for the parameters, $(\mathbf{V}, \boldsymbol{\Theta}_2)$, to be trained. In simulations the accuracy of the refined estimate $\widehat{\mathbf{B}}_{\text{NN}}$ in (3.11) was very sensitive to the initialization of $\mathbf{V}$. We conjecture that a consistent estimator for $\mathbf{B}$ in (1.18), such as (3.9), is required in order to obtain a consistent estimate from the refinement procedure in (3.11).

3.3 Algorithm

In this section the algorithm to obtain the estimates $\widehat{\mathbf{B}}_{\texttt{NN\_OPG}}$ in (3.9) and $\widehat{\mathbf{B}}_{\text{NN}}$ in (3.11) are described. Both estimators depend on training a `MLP` using `tensorflow` [AAB+15] with an `R` interface provided by the `R`-package [AT20] used to implement them.

For training the neuronal networks we use the `RMSProp` [Geo12] algorithm which is a variant of the (mini-batch) *stochastic gradient descent* (SGD) algorithm [Bot98] (see also: [GBC16] and [AT20]).

For regularisation during the training, we apply `dropout` with a rate of 0.4 (see [SHK+14]) after each fully connected hidden layer, i.e. during each update step in the training procedure the nodes are randomly set to 0 with probability 0.4.

For a sample $(Y_i, \mathbf{X}_i^T)_{i=1,\dots,n}^T$ fix natural numbers $m \leq n$, ep where the former is called `batch_size` and the later `number of epochs`. Let $f_{\text{MLP}_N}(\mathbf{x}; \boldsymbol{\Theta})$ be an N-layer `MLP` and the objective function is given by

$$T(\boldsymbol{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(Y_i, f_{\text{MLP}_N}(\mathbf{X}_i; \boldsymbol{\Theta})\right)$$

A rough outline of stochastic gradient descent (`SGD`) is given by

Result: $\widehat{\boldsymbol{\Theta}}^{(\text{end})} = \text{argmin}_{\boldsymbol{\Theta}} T(\boldsymbol{\Theta})$
Initialize: $\boldsymbol{\Theta}^{(0)}$
for $u \in \{1, \ldots, ep\}$ **do**
  for $j \in \{1, \ldots, \lfloor n/m \rfloor\}$ **do**
    Determine the step sizes $\tau \in \mathbb{R}^{\dim(\boldsymbol{\Theta})}$ by `RMSProp`
    $\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} + \text{diag}(\tau) \sum_{l=(j-1)m+1}^{\min(jm,n)} \nabla_{\boldsymbol{\Theta}} \mathcal{L}(Y_l, f_{\text{MLP}_N}(\mathbf{X}_l; \boldsymbol{\Theta}^{(k)}))$
  end
  Shuffle the dataset $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}^T$ randomly
end

Algorithm 2: Stochastic gradient descent outline

If the sample size n is not a multiple of the batch size m, then in the last run of the inner loop the sum of gradients is extended to n. Further if there are restrictions placed on some of the parameters (as is the case for $\mathbf{V} \in \mathcal{S}(p, k)$ in (3.4), the projection back on the Stiefel manifold in (1.1) is done via a standard *Polar decomposition*), the corresponding part of the parameter vector $\boldsymbol{\Theta}$ is projected back to the restricted set after applying the update step.

Further an important feature of stochastic gradient descent training is that the complexity is linear in the sample size n if the number of epochs ep and the batch size m are chosen independently of n. Moreover, for large sample sizes n we observed in simulations that less epochs ep suffice to find well trained neural nets.

The proposed NN estimator for the *mean subspace* is a two stage procedure.

Stage 1: Obtain the $\widehat{\mathbf{B}}_{\texttt{NN\_OPG}}$ estimator in (3.9) by solving the optimization in (3.6). This is done via the stochastic gradient descent (SGD) algorithm with random initialization of the starting value $\boldsymbol{\Theta}_1^{(0)}$ to obtain the estimate $\widehat{\boldsymbol{\Theta}}_1 = (\widehat{\mathbf{W}}_1^{\texttt{NN\_OPG}}, \widehat{\mathbf{b}}_1^{\texttt{NN\_OPG}})$ in (3.6)

Stage 2: Solve the optimization in (3.11) via the stochastic gradient descent (SGD) algorithm. The initial parameters of g_{NN} in (3.4) are set to $(\widehat{\mathbf{B}}_{\texttt{NN\_OPG}}, \boldsymbol{\Theta}_2^{(0)})$ where $\boldsymbol{\Theta}_2^{(0)} = (\widehat{\mathbf{W}}_1^{\texttt{NN\_OPG}} \widehat{\mathbf{B}}_{\texttt{NN\_OPG}}, \widehat{\mathbf{b}}_1^{\texttt{NN\_OPG}})$.

In the second stage, we use the weights and bias obtained by training $g_{\texttt{NN\_OPG}}$ in (3.2) as initialization for the parameters of $g_{\texttt{NN\_wrap}}$ in (3.3) and $\widehat{\mathbf{B}}_{\texttt{NN\_OPG}}$ as initial value for $\mathbf{V} \in \mathcal{S}(p, k)$ in (3.4).

This two stage initialization scheme is important for the performance of the proposed estimator since a random initialization of the parameters of the second stage yielded much worse results.

3.4 Analogy of nn – sdr estimation to mave

The optimization in (1.45) corresponds to local linear smoothing of $\mathbb{E}(Y_i \mid \mathbf{V}^T \mathbf{X}_i)$ with weights given in (1.44). After the local linear estimates $\hat{a}_j, \hat{\mathbf{b}}_j$ in (1.46) are obtained, assume that the weights in (1.44) are given by $w_{i,j}(\mathbf{V}) = 1$ if $i = j$ and 0 if $i \neq j$. Then,

the target function of MAVE in (1.46) can be written as

$$T_n(\mathbf{V}) = \frac{1}{n}\sum_{j=1}^{n}\hat{\sigma}^2(\mathbf{V}^T\mathbf{X}_j) = \frac{1}{n}\sum_i \left(Y_i - \widehat{\mathbb{E}}(Y_i \mid \mathbf{V}^T\mathbf{X}_i)\right)^2 = \frac{1}{n}\sum_i \mathcal{L}\left(Y_i, \widehat{\mathbb{E}}(Y_i \mid \mathbf{V}^T\mathbf{X}_i)\right) \tag{3.12}$$

where $\mathcal{L}$ is the squared error loss and $\widehat{\mathbb{E}}(Y_i \mid \mathbf{V}^T\mathbf{X}_i)$ the local linear smooth. Under this simplifying assumption, (3.12) is the same as (1.39) except that the conditional expectation is estimated via local linear smoothing in MAVE as opposed to neural nets for NN in (3.11).

3.5 Analogy of `NN_OPG` to opg

The OPG estimator estimates $\boldsymbol{\Sigma}_\nabla$ in Theorem 7 via local linear smoothing of $\nabla g(\mathbf{B}^T\mathbf{X}_i)$. Specifically, if $\mathbf{V} = \mathbf{I}_p$ in (1.46), we let $(a_j, \mathbf{b}_j{}^T)_{j=1}^n$ denote the solutions of the optimization in (1.46). Then, $\mathbf{b}_j$ is an estimate for $\nabla g(\mathbf{B}^T\mathbf{X}_j)$ and

$$\widehat{\boldsymbol{\Sigma}}_\nabla = \frac{1}{n}\sum_{j=1}^{n}\mathbf{b}_j\mathbf{b}_j{}^T \tag{3.13}$$

is an estimator for $\boldsymbol{\Sigma}_\nabla$ in Theorem 7.

Definition 19. *The outer product gradient (OPG) estimator for* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \operatorname{span}\{\mathbf{B}\}$ *is defined as*

$$\widehat{\mathbf{B}}_{\text{OPG}} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \tag{3.14}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_k$ *are the first* k *eigenvectors of* (3.13).

3.6 Simulations for nn − sdr

We compare the estimation accuracy of NN estimation with the forward model based sufficient dimension reduction methods, and *mean outer product gradient estimation* (`meanOPG`), *mean minimum average variance estimation* (`meanMAVE` [XTLZ02]) and *conditional variance estimator* (CVE) [FB21a] introduced in Chapter 2. The first two, `meanOPG` and `meanMAVE`, are implemented in the `R`-package [WY19], and CVE in the `R` package `CVarE` [KF21].

We report results for three architectures for $g_{\texttt{NN\_OPG}}$ and $g_{\texttt{NN\_wrap}}$ used in NN estimation. The first is a single layer `MLP` with 128 hidden neurons, the second has 512 and the third is a two layer `MLP` with 48 hidden neurons each. The results were largely undifferentiated for hidden neuron values between 128 and 512. For the two layer MLP, we obtained similar results for more than 48 neurons, which is already a small number. All three architectures use dropout (see [SHK+14]) with probability 0.4[1] after each fully connected hidden layer

[1] Dropout rates ranging from 0 to 0.6 were tried and 0.4 was found to yield the best accuracy in reduction estimation.

except in the reduction layer of the g_{NN}. All architectures in (3.2) are trained in (3.6) with $ep = 200$ epochs and `batch_size` $m = 32$. The refinement training in (3.11) uses $ep = 400$ `epochs` and again `batch size` $m = 32$. We use the estimation algorithm in Section 3.3. The code is available at https://git.art-ist.cc/daniel/NNSDR.

We consider the same six models (M1-M6) as in Section 2.7.2 of Chapter 2, which are reproduced in Table 3.1. We set $p = 20$ throughout. For M1-M5, we let $\mathbf{b}_1 = (1,1,1,1,1,1,0,...,0)^T/\sqrt{6}$, $\mathbf{b}_2 = (1,-1,1,-1,1,-1,0,...,0)^T/\sqrt{6} \in \mathbb{R}^p$. For M6, $\mathbf{b}_1 = \mathbf{e}_1, \mathbf{b}_2 = \mathbf{e}_2$ and $\mathbf{b}_3 = \mathbf{e}_p$, where $\mathbf{e}_j$ denotes the p-vector with jth element equal to 1 and all others are 0. In M7, the first three columns are the identity vectors and $\mathbf{b}_4 = (2\mathbf{e}_4 + \mathbf{e}_5)/\sqrt{5}$ which is taken from [FTAW20]. The error term ϵ is independent of $\mathbf{X}$ for all models. In M2, M3, M4, M5 and M6, $\epsilon \sim N(0,1)$. For M1, ϵ has a generalized normal distribution $GN(a,b,c)$ with densitiy $f_\epsilon(z) = c/(2b\Gamma(1/c))\exp((|z-a|/b)^c)$ [see [Nad05]], with location 0 and shape-parameter 0.5 for M1, and the scale-parameter is chosen such that $\mathbb{V}\text{ar}(\epsilon) = 0.25$. The dimension k is assumed to be known throughout.

Table 3.1: Models

Name	Model	$\mathbf{X}$ distribution	ϵ distribution	k	n
M1	$Y = \cos(\mathbf{b}_1^T\mathbf{X}) + \epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$	$GN(0, \sqrt{1/2}, 0.5)$	1	100
M2	$Y = \cos(\mathbf{b}_1^T\mathbf{X}) + 0.5\epsilon$	$\mathbf{X} \sim Z\mathbf{1}_p + N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	1	100
M3	$Y = 2\log(\lvert\mathbf{b}_1^T\mathbf{X}\rvert + 2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	1	100
M4	$Y = (\mathbf{b}_1^T\mathbf{X})/(0.5 + (1.5 + \mathbf{b}_2^T\mathbf{X})^2) + 0.5\epsilon$	$\mathbf{X} \sim N_p(0, \mathbf{\Sigma})$	$N(0,1)$	2	200
M5	$Y = \cos(\pi\mathbf{b}_1^T\mathbf{X})(\mathbf{b}_2^T\mathbf{X} + 1)^2 + 0.5\epsilon$	$\mathbf{X} \sim U([0,1]^p)$	$N(0,1)$	2	200
M6	$Y = (\mathbf{b}_1^T\mathbf{X})^2 + (\mathbf{b}_2^T\mathbf{X})^2 + (\mathbf{b}_3^T\mathbf{X})^2 + 0.5\epsilon$	$\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{I}_p)$	$N(0,1)$	3	200
M7	$Y = 10\sin(\pi(\mathbf{b}_1{}^T\mathbf{X})(\mathbf{b}_2{}^T\mathbf{X})) + 20(\mathbf{b}_3{}^T\mathbf{X} - 0.5)^2 + 5^{3/2}\mathbf{b}_4{}^T\mathbf{X} + 5\epsilon$	$\mathbf{X} \sim U([0,1]^p)$	$N(0,1)$	4	600

The variance-covariance structure of $\mathbf{X}$ in models M1 and M4 satisfies $\mathbf{\Sigma}_{i,j} = 0.5^{|i-j|}$ for $i,j = 1,\ldots,p$. In M5, $\mathbf{X}$ is uniform with independent entries on the p-dimensional hypercube. The link functions of M4 is studied in [XTLZ02], but we use $p = 20$ instead of 10 and a non identity covariance structure for M4. In M2, $Z \sim 2\text{Bernoulli}(0.3) - 1 \in \{-1, 1\}$, where $\mathbf{1}_q = (1,1,...,1)^T \in \mathbb{R}^q$, this yields that $\mathbf{X}$ has a mixture normal distribution with a mixture probability of 0.3. M7 is a challenging four dimensional model studied in [FTAW20].

As in Chapter 2, we generate $r = 100$ replications of models M1 - M7 and estimate $\mathbf{B}$ using the different sufficient dimension reduction methods. The accuracy of the estimates is assessed using

$$err = \frac{\|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\widehat{\mathbf{B}}}\|}{\sqrt{2k}}, \tag{3.15}$$

which lies in the interval $[0,1]$. The factor $\sqrt{2k}$ normalizes the distance, with values closer to zero indicating better agreement and values closer to one indicating strong disagreement.

We report the average err and their standard deviations in Table 3.2. All three network architectures, $\text{NN}_{128} - \text{SDR}$, NN_{512}, $\text{NN}_{48,48}$ yield similar results, highlighting the robustness of the method with respect to the architecture. We choose NN_{512} as our default setup for the following simulations in Section 3.7. For M1 and M2, CVE yields the most accurate estimation of the reduction $\mathbf{B}$, followed by the $\text{NN} - \text{SDR}$ estimators. OPG and MAVE show the worst performance for the first two models. In M3, the $\text{NN} - \text{SDR}$ estimators are on par

Table 3.2: Mean and standard deviation of estimation errors for M1-M7

Model	OPG	MAVE	CVE	NN_{128}	NN_{512}	$NN_{48,48}$
M1 mean	0.605	0.535	**0.396**	0.450	0.460	0.502
(sd)	(0.179)	(0.207)	(0.108)	(0.126)	(0.152)	(0.200)
M2 mean	0.918	0.910	**0.455**	0.635	0.619	0.752
(sd)	(0.079)	(0.094)	(0.090)	(0.177)	(0.187)	(0.174)
M3 mean	0.754	0.702	0.594	0.608	**0.578**	0.628
(sd)	(0.216)	(0.258)	(0.209)	(0.211)	(0.196)	(0.228)
M4 mean	0.431	0.435	0.572	**0.408**	0.413	0.413
(sd)	(0.095)	(0.099)	(0.131)	(0.088)	(0.082)	(0.073)
M5 mean	**0.415**	0.422	0.441	0.547	0.554	0.601
(sd)	(0.103)	(0.117)	(0.085)	(0.137)	(0.158)	(0.139)
M6 mean	0.181	0.160	0.420	0.133	**0.122**	0.147
(sd)	(0.027)	(0.022)	(0.111)	(0.015)	(0.013)	(0.017)
M7 mean	0.641	**0.637**	0.791	0.698	0.654	0.687
(sd)	(0.074)	(0.071)	(0.032)	(0.051)	(0.074)	(0.068)

Table 3.3: Mean and standard deviation of out of sample-prediction errors for M1-M7

Model	OPG	MAVE	CVE	NN_{128}	NN_{512}	$NN_{48,48}$
M1 mean	0.523	0.427	**0.364**	0.409	0.421	0.422
(sd)	(0.218)	(0.144)	(0.059)	(0.134)	(0.187)	(0.172)
M2 mean	0.736	0.738	**0.396**	0.476	0.506	0.535
(sd)	(0.145)	(0.092)	(0.044)	(0.086)	(0.111)	(0.110)
M3 mean	0.525	0.518	0.432	0.417	0.430	**0.410**
(sd)	(0.110)	(0.107)	(0.083)	(0.092)	(0.089)	(0.088)
M4 mean	0.711	0.713	0.647	**0.438**	0.497	0.470
(sd)	(0.089)	(0.104)	(0.096)	(0.071)	(0.135)	(0.062)
M5 mean	0.462	0.461	**0.440**	0.494	0.482	0.555
(sd)	(0.051)	(0.046)	(0.043)	(0.109)	(0.103)	(0.099)
M6 mean	0.838	0.765	2.354	0.782	**0.612**	1.216
(sd)	(0.177)	(0.228)	(0.914)	(0.117)	(0.081)	(0.224)
M7 mean	33.112	**33.066**	33.884	33.955	35.272	34.136
(sd)	(1.961)	(1.973)	(1.752)	(1.910)	(2.383)	(1.836)

with CVE and OPG, whereas MAVE exhibits the worst performance. In M4, the NN − SDR estimators are on par with OPG, MAVE, while CVE is slightly worse than the rest. In M5, OPG and MAVE are the most accurate, with CVE nearly on par. For M6, the NN − SDR estimators yield the best results followed by OPG and MAVE. M7 is challenging for all methods, with MAVE, OPG, and NN_{512} − SDR the best performing three.

The NN_{512} − SDR estimator is better or on par with OPG, MAVE, and CVE except for M5. This is not surprising in the case of NN − SDR and OPG/MAVE as they are built on a similar idea. The main difference is that MAVE uses local linear smoothing instead of neural nets.

Furthermore, in Table 3.3 we report the mean and standard deviation for the out of sample prediction errors in M1-M7 over $r = 100$ replications. For each data set and replication, we sampled a test set with sample size 1000 from each model and predicted the response Y via the `predict` function in `R` for OPG, MAVE, and CVE. For NN − SDR, the predictions are given by $g_{\text{NN}}(\mathbf{X}_{\text{new}}, (\widehat{\mathbf{B}}_{\text{NN}}, \widehat{\boldsymbol{\Theta}}_1))$ in (3.4). For M1 and M2, CVE gives the smallest out of sample prediction errors, followed by the NN − SDR estimators which outperform both OPG and MAVE. For M3, all three NN − SDR estimators are better or on par with CVE and outperform OPG and MAVE. In M4, NN − SDR outperforms all, with CVE the next best. For M5, CVE performs better than all other. Interestingly, in M6 CVE and $\text{NN}_{48,48}$ − SDR do not work well in terms of prediction accuracy. In M7, MAVE performs the best followed by OPG and CVE, but the NN − SDR estimators trail closely.

In sum, for relatively small to medium samples with few predictors ($p = 20$), NN − SDR exhibits approximately similar and sometimes better performance than its SDR competitors.

3.7 Large sample size simulation for nn − sdr

In this section we simulate data from models M6 and M7 in Section 3.6 and increase both the number of predictors p and the sample size n. We monitor the estimation accuracy by *err* in (3.15) as in Section 3.6, the out of sample prediction error and the required time for the estimation of a reduction.

We examined two simulation settings. In the first, we simulated from model M7 using the same $p = 20$ and increased the sample size significantly ($n = 2^u$, $u = 7, 9, 11, 13$). The results are displayed in Table 3.4. We do not report values for $n = 2048, 8192$ for CVE as the runtime is too long. For $n \in \{128, 512\}$, MAVE is on par with NN_{512} − SDR, whereas for $n = 2048, 8192$, NN_{512} − SDR is slightly more accurate.

To explore how simultaneous growth of the sample size and the number of predictors affect performance, the second simulation revisits M6, where we successively increase both the sample size n and p. The sample sizes considered are $n \in \{1000, 4000, 16000, 64000, 256000\}$ with corresponding $p \in \{32, 63, 126, 253, 506\}$, which is roughly $p \propto \sqrt{n}$. We observed that for larger sample sizes, fewer epochs in the training phase of the neural net suffice. To demonstrate this, the number of epochs was reduced as n and p increased, as follows. For $(n, p) = (1000, 32)$, 200 and 400 epochs were used in the two steps of the refined NN, respectively, and at each subsequent setting, epoch numbers were halved.

The results of this simulation are shown in Table 3.5, which reports the mean and standard deviation (in parentheses) over 10 repetitions of *err* in (3.15), the out of sample prediction errors, and the runtime as measured internally via the user time obtained by the

R function `system.time()`. The advantage of NN emerges in Table 3.5. As both n and p grow, MAVE is no longer computable in realistic time. For example, for $n = 64000, p = 253$, one calculation for MAVE takes about 12 hours to complete. Hence, we report only one value for *err* and prediction error. In contrast, NN takes about 9 minutes to complete one run for the same setting and about 28 minutes to complete one run for $n = 256000, p = 506$. For $n = 1000, 4000, 16000$, and $p = 32, 63, 126$, NN − SDR exhibits slightly higher values of estimation error and lower values of out-of-sample prediction error than MAVE.

The mean runtimes of the two methods are plotted against the sample size in Figure 3.3. We see that the runtime for MAVE explodes to exceed 12 hours only for one dataset at sample size 64000. On the other hand, NN computes in reasonable time.

Thus, NN − SDR is the only forward model based SDR method that is applicable to truly large data while obtaining small estimation and out-of-sample prediction errors. Moreover, for smaller data sets, both in terms of n and p, it maintains competitive performance.

Table 3.4: Mean and standard deviation (in parentheses) of estimation error for model M7

n		OPG	MAVE	CVE	NN_{512}
128	mean	0.802	**0.797**	0.834	0.801
	(sd)	(0.02768)	(0.03561)	(0.02567)	(0.03541)
512	mean	0.691	**0.683**	0.778	0.697
	(sd)	(0.05700)	(0.05923)	(0.03528)	(0.03639)
2048	mean	0.233	0.253		**0.209**
	(sd)	(0.03161)	(0.07841)		(0.06000)
8192	mean	0.102	0.107		**0.082**
	(sd)	(0.00738)	(0.00935)		(0.00722)

3.8 Data Analysis

We analyze three data sets. The first in Section 3.8.1 is of relatively small sample size $n = 506$ and number of predictors $p = 12$, the second in Section 3.8.2 is of large $n = 21613$ and small $p = 16$, and the third in Section 3.8.3 is of very large $n = 382168$ and small to medium $p = 40$.

3.8.1 Boston Housing

In this section we apply the refined NN estimator on the `Boston Housing` data and compare its performance with the other two mean subspace SDR methods, MAVE and CVE. 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 comprise of 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 all three methods operate under the assumption of continuous predictors. The target variable is the median value of owner-occupied homes, `medv`, in $1,000. The 12 predictors

Table 3.5: Mean and standard deviation (in parentheses) of *err*, out of sample prediction error, and runtime for model M6.

n	p	Method	*err*	MPE	time [sec]
1000	32	MAVE	0.0626	0.393	5.48
			(0.003)	(0.028)	(0.031)
		NN_{512}	0.0547	0.343	48.65
			(0.004)	(0.021)	(0.700)
4000	63	MAVE	0.0445	0.351	71.20
			(0.002)	(0.019)	(0.842)
		NN_{512}	0.0496	0.313	91.35
			(0.003)	(0.016)	(0.822)
16000	126	MAVE	0.0323	0.337	1416.14
			(0.001)	(0.016)	(34.367)
		NN_{512}	0.0631	0.329	215.78
			(0.002)	(0.025)	(1.793)
64000	253	MAVE	0.023	0.325	$\sim 12h$
			(0) *	(0) *	(0) *
		NN_{512}	0.095	0.387	542.26
			(0.001)	(0.019)	(2.934)
256000	506	NN_{512}	0.153	0.568	1673.03
			(0.003)	(0.028)	(6.650)

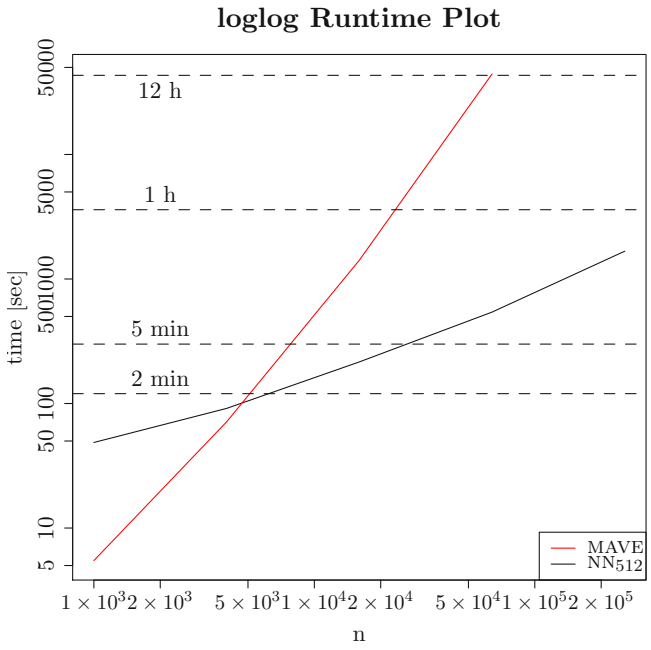
* Only one repetition was run as it takes about 12 hours.

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 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 set the dimension of the reduction $\mathbf{B}$ to two; i.e., $k = 2$, for all three methods and compute prediction errors using squared error loss and leave-one-out cross validation. The NN with one layer and 512 neurons is fitted on the $n-1$ training data and compute the predicted value for the left out data point. Both CVE and MAVE were applied to the standardized training data. The mean and standard deviation (in parentheses) of the 506 prediction errors are displayed in Table 3.6 [2]. The CVE method results in the smallest prediction error followed by $NN-SDR$, which, on the other hand, has the smallest standard error. MAVE is the least accurate. The analysis for $k = 1$ yielded similar results. In this

[2] In Table 3.6 the out of sample prediction errors are non standardized (the reduction is estimated from the standardized training set but the predictions of the response are on the original scale) in contrast to the values reported in Table 2.4 in Chapter 2

Figure 3.3: Runtime comparison of MAVE against NN_{512} with equivalent estimation performance.



example of small n-small p, nonparametric methods are expected to do well, which is what we observe for CVE followed by MAVE. Nevertheless, the performance of the large sample NN − SDR method is roughly on par with both.

Table 3.6: Leave-One-Out Cross Validation Prediction errors with reduction dimension $k = 2$.

	MAVE	CVE	NN_{512}
mean	18.762	16.148	18.006
(sd)	(63.136)	(63.500)	(41.739)

3.8.2 KC Housing

Further, we compare NN estimation to MAVE using the `kc_house_data` set in the `R` package MAVE. The data set contains 21613 observations on 20 variables. The target variable is `price`, the price of a sold house. We use 16 predictors after omitting `id, date`, and `zip code`: `bedrooms` (number of bedrooms), `bathrooms` (number of bathrooms), `sqft_living` (square footage of the living room(, `sqrt_log` (square footage of the log), `floors` (total floors in the house), `waterfront` (whether the house has a view a waterfront(1: yes, 0: not)), `view` (unknown), `condtion` (condition of the house), `grade` (unknown), `sqft_above` (square footage of house apart from basement), `sqft_basement` (square footage of the

basement), `yr_built` (built year), `yr_renovated` (year when the house was renovated), `lat` (latitude coordinate), `long` (longitude coordinate), `sqft_living15` (living room area in 2015(implies some renovations)), `sqrt_lot15` (lot area in 2015(implies some renovations)).

We perform 10-fold cross-validation in order to obtain an unbiased estimate of the out of sample prediction error. We set $k = 1$ and report the average fraction of the mean squared prediction error divided by the variance of the response on the test set, as well as its standard error, in Table 3.7. Our NN − SDR estimator has out of sample mean squared error that is about half the variance of the response on the test set, whereas MAVE's is less than 2 percent lower than the variance of the response. This means that the NN − SDR regression explains roughly half of the total variance in the response whereas MAVE hardly explains any. Further, even though the MAVE reduction is estimated in roughly the same time as NN − SDR, in 6 out of the 10 folds the `predict` function for MAVE produces an error. We also report the 10-fold cross-validated prediction error for CVE, which yields the best result as it explains more than 70% of the total variance in the response but could not be computed, in its current implementation, on a personal computer. [3]

The coefficients of the reductions are given in Table 3.8. NN − SDR extracts information from all variables as it places non-zero weights of varying size on all. MAVE, on the other hand, selects `waterfront` and the co-linear `sqft_living, sqft_above, sqft_basement` (`sqft_living` = `sqft_above` + `sqft_basement`) and drops all other variables. Moreover, it allocates the same weight to the collinear variables with opposite signs, effectively discounting all three and ultimately declaring only `waterfront` relevant.

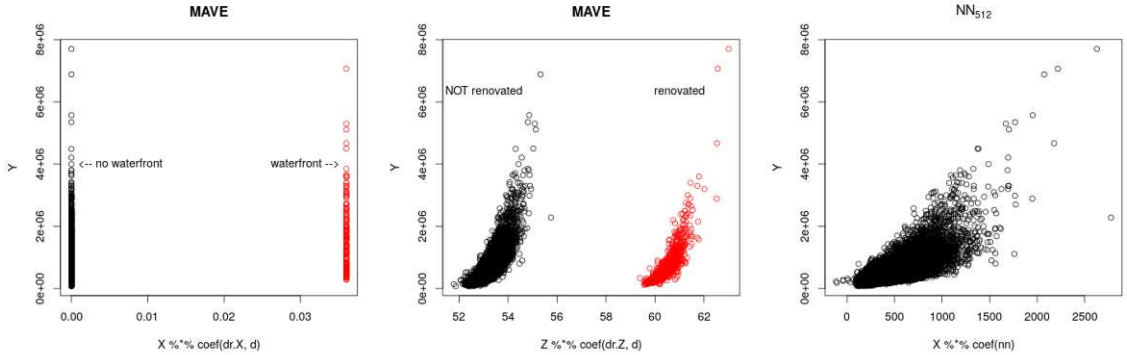
These results indicate that MAVE breaks down in the analysis of this data set. Since `sqft_living` = `sqft_above` + `sqft_basement`, we dropped `sqft_basement` to investigate the effect of collinearity. In Figure 3.4, we plot the response versus the MAVE reduction computed on all predictors in the left panel, versus the MAVE reduction without `sqft_basement` and versus the NN − SDR reduction in the right panel. The reduced predictors are strikingly different. The NN − SDR reduction is smooth and captures a clear nonlinear heteroskedastic relationship with `price`. The plot in the left panel captures the failure of MAVE to extract the predictive information in the predictors, as no apparent pattern emerges. Moreover, the data are arbitrarily split in the groups defined by the binary `waterfront` variable. Once the collinearity is removed, MAVE captures the relationship between Y and $\mathbf{X}$ but nevertheless it again splits the data into two new arbitrary classes for the renovated and non-renovated houses. This variable takes either value 0 (not renovated) or the renovation year that ranges between 1934 and 2015. The black points in the middle panel correspond to 0 and red to the period 1934-2015.

We further draw attention to the semblance of the data clouds across the two categories in the middle panel and the NN − SDR reduction in the right panel. Both MAVE and NN − SDR discover the same pattern, with the correlation coefficients of MAVE and NN − SDR reductions being 0.82 and 0.85, albeit MAVE introduces an artificial split in the data.

In Table 3.8, we also provide the coefficients of the last two eigenvectors, corresponding to the two smallest eigenvalues in decreasing order, of the sample covariance matrix of the predictors. The next to last places most of the weight on `waterfront` and the last on `sqft_living` and `sqft_above, sqft_basement`. Moreover, the vector of coefficients of

[3] The CVE values were computed on the *Vienna Scientific Cluster* (VSC).

Figure 3.4: Reduced data versus response of the `kc_house_data` for MAVE and NN_{512}.



the MAVE reduction based on all predictors in the first column seems to be the sum of the last and the down-weighted second to last eigenvectors of the sample covariance matrix of $\mathbf{X}$. This relates to the fact that the sample covariance matrix of $\mathbf{X}$ is singular of rank $16 = p - 1$. Thus, the last eigenvector dominates all others and largely agrees with the MAVE reduction coefficients. We investigate the effect of collinearity on MAVE and CVE in Section 3.8.2.1.

Table 3.7: Ten-fold Cross Validation Relative Prediction errors with reduction dimension $k = 1$.

	MAVE	CVE	NN_{512}
mean	0.982	0.296	0.527
(sd)	(0.035)	(0.149)	(0.043)

3.8.2.1 The case of singular $\mathbf{\Sigma_X}$

We consider the effect of collinear predictors on the sufficient dimension reduction techniques MAVE, CVE, and NN − SDR. We assume that $\mathbf{\Sigma_X} = \mathbb{V}\mathrm{ar}(\mathbf{X})$ is singular and show that, in this case, the mean subspace is not uniquely identifiable [4].

Let $\mathbf{U}$ be a basis of the nullspace of $\mathbf{\Sigma_X}$, consisting of the eigenvectors that correspond to the 0 eigenvalue. Without loss of generality, we assume the eigenspace of $\mathbf{U}$ to be one dimensional. Then $\mathbf{U}^T\mathbf{X} = c$ is constant and we can write

$$Y = g(\mathbf{B}^T\mathbf{X} + c - c) + \epsilon = g_c((\mathbf{B} + \mathbf{U})^T\mathbf{X}) + \epsilon \tag{3.16}$$

$$= g_c(\tilde{\mathbf{B}}^T\mathbf{X}) + \epsilon \tag{3.17}$$

where $g_c(\mathbf{x}) = g(\mathbf{x} - c)$ fulfills all assumptions of the link function in model (1.18) and $\tilde{\mathbf{B}} = \mathbf{B} + \mathbf{U}$. If $\mathrm{span}\{\mathbf{U}\} \subset \mathrm{span}\{\mathbf{B}\}$, then $\mathrm{span}\{\mathbf{B}\} = \mathrm{span}\{\tilde{\mathbf{B}}\}$ and the mean subspace

[4] In this case the density $f_{\mathbf{X}}$ does not have a convex support on $\mathbb{R}^p$ which is required in assumption (A.1) in Section 2 and guarantees the existence and uniqueness of the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ via proposition 6.4 in [Coo98c]

Table 3.8: Estimated linear reductions for the `kc_house_data` data set. $\mathbf{Z}$ are the remaining 16 predictors when dropping `sqft_basement` from $\mathbf{X}$. The last column $v_{17}(\mathbf{X}), v_{16}(\mathbf{X})$ are the PCA coefficients corresponding to the last two smallest eigenvalue of $\mathbb{V}\text{ar}(\mathbf{X})$ for un-scaled $\mathbf{X}$.

	$\hat{\mathbf{B}}_{\text{MAVE}(\mathbf{X})}$	$\hat{\mathbf{B}}_{\text{MAVE}(\mathbf{Z})}$	$\hat{\mathbf{B}}_{\text{NN}_{512}(\mathbf{X})}$	$v_{17}(\mathbf{X})$	$v_{16}(\mathbf{X})$
bedrooms	0.000	-0.015	-0.171	0.000	-0.005
bathrooms	0.000	0.050	0.057	0.000	-0.001
sqft_living	0.577	0.000	0.099	0.577	0.000
sqft_lot	0.000	0.000	0.000	0.000	0.000
floors	0.000	0.036	0.098	0.000	0.000
waterfront	0.036	0.330	0.485	0.000	-0.999
view	0.000	0.047	0.787	0.000	0.046
condition	0.000	0.042	0.152	0.000	0.002
grade	0.000	0.124	0.204	0.000	-0.003
sqft_above	-0.577	0.000	0.080	-0.577	0.000
sqft_basement	-0.577		0.112	-0.577	0.000
yr_built	0.000	-0.003	-0.031	0.000	0.000
yr_renovated	0.000	0.004	0.051	0.000	0.000
lat	0.000	0.925	-0.001	0.000	-0.011
long	0.000	-0.107	0.007	0.000	-0.016
sqft_living15	0.000	0.000	0.079	0.000	0.000
sqft_lot15	0.000	0.000	-0.001	0.000	0.000

is unique. Otherwise, both span$\{\mathbf{B}\}$ and span$\{\tilde{\mathbf{B}}\}$ are dimension reduction subspaces but span$\{\tilde{\mathbf{B}}\} \neq$ span$\{\mathbf{B}\}$.

Most SDR approaches, including MAVE [XTLZ02, Cond. 3(a), p. 386] and CVE [FB21a, Cond. A.1 , p. 3] require $\mathbf{X}$ have a density with convex support; that is, its variance-covariance is positive definite. It appears that MAVE is more sensitive to the violation of this assumption as compared to CVE.

To demonstrate this we present a small simulation study in the presence of near collinearity. Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ with $(X_2, \ldots, X_p) \sim N(\mathbf{0}, \mathbf{I}_{p-1})$ and $X_1 = -0.5(X_2 + X_3) + 0.001Z$, where $Z \sim N(0,1)$ and is independent of $(X_2, \ldots, X_p)$. Then, $\mathbf{U} = (2,1,1,0,\ldots,0)/\sqrt{6} \approx (0.816, 0.408, 0.408, 0, \ldots, 0)$ is the eigenvector of $\mathbf{\Sigma_x}$ corresponding to the smallest eigenvalue.

Let $p = 10$ and $Y = (\mathbf{B}^T\mathbf{X})^2 + 0.5\epsilon$, where $\epsilon \sim N(0,1)$ is independent from $\mathbf{X}$ and $\mathbf{B} = \mathbf{e}_4$ the fourth standard basis vector. We draw 100 random samples $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}^T$ of size $n = 100$ from this model and calculate the MAVE, CVE and NN_{512} estimators of $\mathbf{B}$. The median, mean and standard deviation of the estimation errors for the subspace in (3.15) are reported in Table 3.9.

For example, one of the $\mathbf{B}_{\text{MAVE}}$ estimates is (0.815, 0.404, 0.402, −0.102, 0,0, −0.001, −0.006, 0.003, −0.008) with associated error 0.995. We can clearly see that MAVE estimates $\mathbf{U}$ instead of $\mathbf{B}$, and most MAVE estimates follow the same pattern. On the other hand,

Table 3.9: Ten-fold Cross Validation Relative Prediction errors mean and standard deviation (in brackets) with strong collinearity in the predictors.

	MAVE	CVE	NN_{512}
mean	0.917	0.164	0.101
median	0.999	0.162	0.096
(sd)	(0.256)	(0.057)	(0.032)

one of the $\mathbf{B}_{\text{CVE}}$ estimates is (0.013, −0.007, −0.018, −0.99, 0.015, −0.004, −0.046, −0.13, 0.015, −0.021), with associated error 0.143 and one of the $\mathbf{B}_{\text{NN}_{512}}$ estimates is (−0.007, −0.002, −0.074, −0.995, 0.013, 0.004, −0.035, −0.047, 0.026, −0.031), with associated error 0.103. CVE and NN − SDR stays clear of $\mathbf{U}$ and correctly identifies the true $\mathbf{B}$.

In this example, in particular, MAVE seems to focus solely on estimating $\mathbf{U}$ instead of $\mathbf{B}$. This does not hold in general. We offer an explanation by setting $c = \alpha c$ in (3.16) for a scalar α. Following the rationale below (3.16), $\tilde{\mathbf{B}} = \mathbf{B} + \alpha\mathbf{U}$ is a reduction for any α. Since MAVE, CVE and NN − SDR work with $\tilde{\mathbf{B}} \in \mathcal{S}(p, k)$, α determines the weight placed on $\mathbf{U}$ relative to $\mathbf{B}$. For large α, $\mathbf{U}$ dominates the reduction $\tilde{\mathbf{B}}$ and MAVE fails to identify the mean subspace. In contrast, CVE and NN − SDR remains robust in its ability to accurately estimate the reduction $\mathbf{B}$.

We conjecture that MAVE's vulnerability is numerical in nature and relates to the implementation algorithm in the MAVE package since the solution of the least square problem used in MAVE is not unique anymore if $\mathbb{V}\text{ar}(\mathbf{X})$ is singular. We also conjecture that CVE and NN − SDR are more robust than MAVE.

3.8.3 Beijing Air Quality Data

The *Beijing Multi-Site Air-Quality Data* [ZGD+17] available at the UCI machine learning repository[5] includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites in Beijing. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration. After removing missing data entries, the data contains 382168 complete measurements.

The target $PM_{2.5}[ug/m^3]$ is the concentration of particle matter in the air with less than 2.5 micrometres in diameter.

The predictors are `year`, `month`, `day`, `hour`, `SO2` (SO2 concentration $[ug/m^3]$), `NO2` (NO2 concentration $[ug/m^3]$), `CO` (CO concentration $[ug/m^3]$), `O3` (O3 concentration $[ug/m^3]$), `TEMP` (temperature $[C^\circ]$), `PRES` (pressure $[hPa]$), `DEWP` (dew point temperature $[C^\circ]$), `RAIN` (precipitation $[mm]$), `wd` (wind direction), `WSPM` (wind speed $[m/s]$), `station` (name of the air-quality monitoring site). The two categorical variables `wd` and `station`, with 16 and 12 categories, respectively, are converted to 26 dummy variables, resulting in 40 predictors.

We included the categorical variables to demonstrate that NN − SDR can handle dummy variables even though it is not designed for this. Given the large sample size we used 2

[5] https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data

epochs for the first stage and 3 for the second refinement stage of the training. Due to the large sample size MAVE and CVE are infeasible to compute while NN − SDR executes in less than 3 minutes per fold run on the CPU of personal computer. As a comparison, we included the *linear model* (`lm`) as well as the *Multivariate Adaptive Regression Splines* (`mars`, [Fri91, HR17]), as both can be applied to large regressions, provided $p < n$ and are computationally efficient.

In Table 3.10, the mean of the 10-fold cross validation prediction errors is reported. The linear model exhibits the worst performance, as expected. NN − SDR improves upon the linear model for all choices of dimension we examined, beats `mars` for $k = 3, 4$ and obtains the minimum MSPE for $k = 4$. Thus, not only is NN − SDR the best method with respect to predictive accuracy, but it also provides an assessment of the true structural dimension of the relationship ($k = 4$) between the response and the predictors. This confirms the improved performance of `mars`, a multivariate nonparametric fitting method, over the linear model and points to the nonlinearity of the relationship.

Table 3.10: 10-Fold Cross Validation Mean Squared Prediction errors.

	`lm`	`mars`	NN_{512} − SDR $k = 1$	NN_{512} − SDR $k = 2$	NN_{512} − SDR $k = 3$	NN_{512} − SDR $k = 4$
mean	1829	1628	1746	1654	1604	1526
(sd)	(20.8)	(24.9)	(19.9)	(18.8)	(24.0)	(59.3)

4 Ensemble Conditional Variance Estimation

In this chapter we introduce the extension of CVE to the exhaustive *ensemble conditional variance estimator* ECVE for the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ (see section 1.4), which is a superset of $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$, i.e.

$$\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}}$$

Since the introduction of the concept of central mean subspace in [CL02], several papers made gradual contributions to establish a road path from the central mean to the central subspace [see [YL11] for a list]. [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, then one can recover the central subspace. They then proposed and studied families of functions, which they called *ensembles*, that were rich enough to obtain the desired outcome.

Throughout this chapter we assume model (1.13). The idea is to apply CVE to identify the mean subspaces of transformed responses $f(Y)$, with f function elements of an *ensemble*, and then combine them to form the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$.

4.1 Ensembles

[YL11] introduced *ensembles* as a device to extend mean subspace SDR methods to central subspace ones. The *ensemble* approach of combining mean subspaces to span the central subspace comprises of two components: (a) a rich family of functions of transformations for the response and (b) a sampling mechanism for drawing the functions from the ensemble to ascertain coverage of the central subspace. To distinguish between families of functions and ensembles, [YL11] use the term *parametric* ensemble, which we also use herein.

Definition 20. *Let $\mathcal{F}$ be a family of measurable functions from $supp(Y) \subset \mathbb{R}$ to $\mathbb{R}$, then $\mathcal{F}$ is called an ensemble. If $\mathcal{F}$ is a parametric family of measurable functions which is measurable with respect to the index, i.e. $\mathcal{F} = \{f_t : t \in \Omega_T\}$ for some index set Ω_T, $\mathcal{F}$ is called parametric ensemble.*

Let $\mathcal{F}$ be an ensemble and set $f(Y)$ where Y is given by model (1.13). We call $\mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}$ the mean subspace of the transformed random variable $f(Y)$, as defined in (1.21), i.e.

$$\mathbb{E}\left(f(Y) \mid \mathbf{X}\right) = \mathbb{E}(f(Y)|\mathbf{P}_{\mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}}\mathbf{X})$$

An ensemble $\mathcal{F}$ must be rich enough to be useful, therefore, following [YL11], we define.

Definition 21. *An ensemble $\mathcal{F}$ characterises $\mathcal{S}_{Y|\mathbf{X}}$, if*

$$\operatorname{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\} = \mathcal{S}_{Y|\mathbf{X}} \tag{4.1}$$

As an example of an ensemble $\mathcal{F}$ that can characterise the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$, we consider the parametric ensemble $\mathcal{F} = \{f_t : t \in \Omega_T\} = \{\exp(it\cdot) : t \in \mathbb{R}\}$, i.e. $f_t(\cdot) = \exp(it\cdot)$ and $\Omega_T = \mathbb{R}$. Then

$$\mathbb{E}(f_t(Y) \mid \mathbf{X}) = \mathbb{E}\left(\exp(itY) \mid \mathbf{X}\right).$$

is the conditional characteristic function evaluated at t. By varying over the parametric ensemble $\mathcal{F}$, i.e. over $t \in \mathbb{R}$, we analyze the whole conditional characteristic function. The characteristic function determines a distribution uniquely and we expect to fully recover the conditional distribution of $Y \mid \mathbf{X}$ and therefore also the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$, i.e.

$$\operatorname{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\} = \operatorname{span}\{\mathcal{S}_{\mathbb{E}(\exp(itY)|\mathbf{X})} : t \in \mathbb{R}\} = \mathcal{S}_{Y|\mathbf{X}}$$

[YL11] provide a list of parametric ensembles $\mathcal{F}$, that can characterize $\mathcal{S}_{Y|\mathbf{X}}$ under some mild regularity assumptions.

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

Indicator ensemble $\mathcal{F} = \{1_{\{z \le t\}} : t \in \mathbb{R}\}$, i.e. $\operatorname{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\}$ describes the conditional cumulative distribution function (cdf)

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

Polynomial ensemble $\mathcal{F} = \{z^t : t = 1, 2, 3, ...\}$, i.e. $\operatorname{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : f_t \in \mathcal{F}\}$ describes 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 intuition is as follows: the *characteristic* and the *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 the response Y exists, then the *kernel* ensemble can characterise the conditional distribution $Y \mid \mathbf{X}$. Further if the conditional moment generating function, which is determined by all conditional moments, exists then the polynomial ensemble can characterise $\mathcal{S}_{Y|\mathbf{X}}$.

Theorem 17, which is Theorem 2.1 of [YL11], establishes when an ensemble $\mathcal{F}$ is rich enough to characterise $\mathcal{S}_{Y|\mathbf{X}}$. Let $\mathcal{B} = \{1_A : A \quad \text{is a borel set in} \quad \operatorname{supp}(Y)\}$ be the set of indicator function on $\operatorname{supp}(Y)$ and $L^2(F_Y) = \{f(Y) : \mathbb{E}(f(Y)^2) < \infty\}$ be the set of square integrable random variables with respect to the distribution F_Y of the response Y, that are measurable with respect to the sigma field generated by Y.

Theorem 17. *If $\mathcal{F} \subseteq L^2(F_Y)$ is dense in $\mathcal{B} \subseteq L^2(F_Y)$ then the ensemble $\mathcal{F}$ can characterises the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.*

[YL11] used the ensemble device to extended *minimum average variance estimation* (MAVE), which targets the mean subspace, to its ensemble version that estimates the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ consistently.

Theorem 18 shows that finitely many functions of an ensemble $\mathcal{F}$ are sufficient to characterise the *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$. The caveat is that we do not now which finite subset of $\mathcal{F}$ is sufficient.

Theorem 18. *If a parametric ensemble $\mathcal{F}$ characterises $\mathcal{S}_{Y|\mathbf{X}}$, then there exist finitely many functions $f_t \in \mathcal{F}$ with $t = 1, \ldots, m$ and $m \in \mathbb{N}$ such that*

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

Proof: Let $k = \dim(\mathcal{S}_{Y|\mathbf{X}}) \leq p$ and note that $\dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}) = k_t \leq k$ since $\mathcal{F}$ characterises $\mathcal{S}_{Y|\mathbf{X}}$. If $k_t = 0$ then we can leave out $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \{\mathbf{0}\}$ in (4.1). If $k_t \neq 0$, then $k_t \geq 1$ and $\{\mathbf{0}\} \subset \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathbb{R}^p$ is at least a one dimensional linear subspace. Then if there are infinitely many $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \neq \{\mathbf{0}\}$ (i.e. $k_t \neq 0$) and they span $\mathcal{S}_{Y|\mathbf{X}}$ then infinitely many $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ must span the same subspace, otherwise we obtain the contradiction

$$\dim\left(\operatorname{span}\{\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} : t \in \Omega_T\}\right) > k = \dim(\mathcal{S}_{Y|\mathbf{X}})$$

□

4.2 Motivation of ecve

Throughout this chapter we refer to the following assumptions as needed.

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

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

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

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

(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 $f \in \mathcal{F}$

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

Remark. *Assumptions (E.1) to (E.4) relate to (A.1) to (A.4) in Section 2 as follows.*

- *Assumption (E.1) is analogous to (A.1) and assures that* $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$ *exists and is unique. Furthermore, it allows the mean subspace to be a proper subset of the central subspace, i.e.* $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subsetneq \mathcal{S}_{Y|\mathbf{X}}$.
- *Assumption (E.2) is analogous to (A.2)*
- *Assumption (E.3) is a generic assumption since it states that the set of indices that characterise 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.*
- *(E.4) replaces (A.2) that the link function* g *in model (1.18) is twice continuously differentiable and (A.3) that the response* Y *has finite 8th moment. If the characteristic or indicator ensemble are used (E.4) states that the conditional characteristic or distribution function are twice continuously differentiable. The existence of the 8th moments are automatically fulfilled due to the boundedness of the complex exponential and indicator function.*

Definition 22. *For* $q \leq p \in \mathbb{N}$, $f \in \mathcal{F}$, *and any* $\mathbf{V} \in S(p,q)$, *we define*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\text{ar}\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\}\right) \tag{4.2}$$

where $\mathbf{s}_0 \in supp(f_{\mathbf{X}})$ *is a shifting point.*

Definition 23. *Let* $\mathcal{F}$ *be a parametric ensemble and* F_T *a probability measure on the index set* Ω_T. *For* $q \leq p$, *and any* $\mathbf{V} \in S(p,q)$, *we define*

$$\begin{aligned} L_{\mathcal{F}}(\mathbf{V}) &= \int_{\Omega_T} \int_{\mathbb{R}^p} \tilde{L}(\mathbf{V}, \mathbf{x}, f_t) dF_{\mathbf{X}}(\mathbf{x}) dF_T(t) \\ &= \mathbb{E}_{t\sim F_T}\left(\mathbb{E}_{\mathbf{X}}\left(\tilde{L}(\mathbf{V}, \mathbf{X}, f_t)\right)\right) = \mathbb{E}_{t\sim F_T}(L^*(\mathbf{V}, f_t)), \end{aligned} \tag{4.3}$$

where $F_{\mathbf{X}}$ *is the cumulative distribution function (cdf) of* $\mathbf{X}$, *with*

$$L^*(\mathbf{V}, f_t) = \mathbb{E}_{\mathbf{X}}\left(\tilde{L}(\mathbf{V}, \mathbf{X}, f_t)\right). \tag{4.4}$$

For the identity function, $f_{t_0}(z) = z$, (4.4) is the target function of *conditional variance estimation* given in (2.2) and (4.2) is the same as (2.1). Further if the random variable t is concentrated on one point t_0, i.e. $t \sim \delta_{t_0}$, that corresponds to the identity function $f_{t_0}(z) = z$, then the *ensemble conditional variance estimator* (ECVE) coincides with the *conditional variance estimator* (CVE).

Furthermore, the following holds:

Theorem 19. *Assume (E.1) holds. Let* $\widetilde{\mathbf{B}}$ *be a basis of* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$, *i.e.* $\text{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$, *then for any* $f \in \mathcal{F}$ *for which assumption (E.4) holds,*

$$f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon} \tag{4.5}$$

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

Theorem 19 obtains that (4.5) has the same form as the forward model (1.18) assumed in CVE with the difference that only $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and not independence of $\mathbf{X}$ and the error $\tilde{\epsilon}$ is required.

Proof of Theorem 19.

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

By the tower property of the conditional expectation it holds $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = \mathbb{E}(f(Y) \mid \mathbf{X}) - \mathbb{E}(\mathbb{E}(f(Y) \mid \mathbf{X}) \mid \mathbf{X}) = \mathbb{E}(f(Y) \mid \mathbf{X}) - \mathbb{E}(f(Y) \mid \mathbf{X}) = \mathbf{0}$. The differentiability follows directly from (E.4). □

Theorem 20. *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 (1.1). Let f such that the requirements in assumption (E.4) hold and $\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x})$, where $\tilde{\epsilon}$ is given in (4.5), is continuous. Then,*

$$\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2 + \tilde{h}(\mathbf{V}, \mathbf{s}_0, f) \tag{4.6}$$

where

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

with $f(Y) = g(\tilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}$ decomposed as in (4.5), $t^{(l)}(\mathbf{V}, \mathbf{s}_0, f)$ of the decomposed response $f(Y)$ is defined in (2.6)[1], and

$$\tilde{h}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\})$$
$$= \int_{supp(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1 / \int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}. \tag{4.7}$$

Further,

$$\mathbf{V}_q^t = \mathrm{argmin}_{\mathbf{V} \in \mathcal{S}(p,q)} L^*(\mathbf{V}, f_t) \tag{4.8}$$

is well defined, where $L^*(\mathbf{V}, f_t)$ given in (4.4) is well defined and continuous, and

$$\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \mathrm{span}\{\mathbf{V}_q^t\}^{\perp}, \tag{4.9}$$

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

[1] For $l = 0$ we have that $t^{(0)}(\mathbf{V}, \mathbf{s}_0) = \int_{\mathbb{R}^q} g(\widetilde{\mathbf{B}}^T\mathbf{s}_0 + \widetilde{\mathbf{B}}^T\mathbf{V}\mathbf{r}_1)^0 f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1 = \int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1$ does not depend on f

Proof of Theorem 20. By assumption (E.1) it holds $Y = g_{cs}(\mathbf{B}^T\mathbf{X}, \epsilon)$ 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 $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$, i.e. $\operatorname{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \operatorname{span}\{\mathbf{B}\}$. Then, by Theorem 19,

$$f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon} \tag{4.10}$$

with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$ and g twice continuously differentiable. Then

$$\begin{aligned}
&\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \mathbb{V}\mathrm{ar}\left(f(Y) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&= \mathbb{V}\mathrm{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) + 2\mathrm{cov}\left(\tilde{\epsilon}, g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&\qquad + \mathbb{V}\mathrm{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) \\
&= \mathbb{V}\mathrm{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) + \mathbb{V}\mathrm{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right)
\end{aligned} \tag{4.11}$$

The covariance term in (4.11) vanishes since the sigma field generated by $\mathbf{X}$ is larger than the one generated by $\mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}$, then the tower property of conditional expectation yields

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

The first term in (4.11) can be handled as in Theorem 8, i.e. $\mathbb{V}\mathrm{ar}\left(g(\widetilde{\mathbf{B}}^T\mathbf{X}) \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right) = \mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2$ is well defined and continuous and attains the minimum of 0 for all $\mathbf{s}_0 \in \operatorname{supp}(f_{\mathbf{X}})$ if $\mathbf{V} \perp \widetilde{\mathbf{B}}$. To obtain (4.6) and (4.7) , let $\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x}) = h(\mathbf{x})$. Then,

$$\begin{aligned}
\mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}) &= \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}) \\
&= \mathbb{E}(\underbrace{\mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X})}_{=h(\mathbf{X})} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}) \\
&= \int_{\operatorname{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1 / t^{(0)}(\mathbf{V}, \mathbf{s}_0, f) = \tilde{h}(\mathbf{V}, \mathbf{s}_0, f)
\end{aligned} \tag{4.12}$$

where the first equality is due to $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$, the second follows from the tower property of the conditional expectation, the third from Theorem 8 (a), and the last is a defining equation of $\tilde{h}(\mathbf{V}, \mathbf{s}_0, f)$. Therefore (4.4) is well defined and continuous by an analogously argument as in the proof of Theorem 8. Moreover, (4.8) exists as the minimizer of a continuous function over the compact set $\mathcal{S}(p, q)$.

Then

$$L^*(\mathbf{V}, f) = \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}}\left(\mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2\right) + \mathbb{E}_{\mathbf{s}_0 \sim \mathbf{X}}\left(\mathbb{V}\mathrm{ar}\left(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \operatorname{span}\{\mathbf{V}\}\right)\right) \tag{4.13}$$

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 (4.13) is constant with respect to $\mathbf{V}$.

By the law of total variance,

$$\begin{aligned}\mathbb{V}\text{ar}(\tilde{\epsilon}) = \mathbb{E}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) + \mathbb{V}\text{ar}\left(\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) \iff \\ \mathbb{V}\text{ar}(\tilde{\epsilon}) = \mathbb{E}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) \qquad (4.14)\end{aligned}$$

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 (4.14) into (4.13) obtains

$$L^*(\mathbf{V}, f_t) = \mathbb{E}\left(\mu_2(\mathbf{V}, \mathbf{X}, f_t) - \mu_1(\mathbf{V}, \mathbf{X}, f_t)^2\right) + \mathbb{V}\text{ar}(\tilde{\epsilon}) \qquad (4.15)$$

where $\mathbb{E}\left(\mu_2(\mathbf{V}, \mathbf{X}, f_t) - \mu_1(\mathbf{V}, \mathbf{X}, f_t)^2\right)$ is as in Theorem 9. Therefore, it is well defined, continuous and attains the minimum of 0 for $\mathbf{V} \perp \widetilde{\mathbf{B}}$. □

Remark. *For a more detailed notation for obtaining (4.14), let $\tilde{\mathbf{X}}$ be an independent copy of $\mathbf{X}$. Then the vector $(\mathbf{X}^T, \tilde{\mathbf{X}}^T, \epsilon)^T \in \mathbb{R}^{2p+1}$ drives all the stochasticity in (4.14) and*

$$\mathbb{E}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \text{span}\{\mathbf{V}\})\right) = \mathbb{E}_{\tilde{\mathbf{X}}}\left(\mathbb{V}\text{ar}(\tilde{\epsilon} | \mathbf{U}^T(\mathbf{X} - \tilde{\mathbf{X}}) = \mathbf{0}, \tilde{\mathbf{X}} = \mathbf{s}_0\right)$$

where $\mathbf{U} \perp \mathbf{V}$.

In Chapter 2, model (1.18) was assumed; i.e., $Y = g(\mathbf{B}^T\mathbf{X}) + \epsilon$ with $\epsilon \perp\!\!\!\perp \mathbf{X}$, which implies $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{B}\} = \mathcal{S}_{Y|\mathbf{X}}$. There we showed that the *conditional variance estimator* (CVE) can identify $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ at the population level.

Theorem 20 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 centered and not necessarily independent from $\mathbf{X}$. This allows CVE to apply to problems where the *mean subspace* is a proper subset of the *central subspace*, i.e. $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} \subsetneq \mathcal{S}_{Y|\mathbf{X}}$.

Next we define the *ensemble conditional variance estimator* (ECVE) for a parametric ensemble $\mathcal{F}$ which characterises 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 (4.3) that indexes the ensemble $\mathcal{F}$ as [YL11].

Definition 21. *The estimation equation of **Ensemble Conditional Variance Estimator** with respect to the ensemble $\mathcal{F}$ on the population level is any basis $\mathbf{B}_{p-q,\mathcal{F}}$ of $\text{span}\{\mathbf{V}_q\}^\perp$, where*

$$\mathbf{V}_q = \text{argmin}_{\mathbf{V} \in S(p,q)} L_{\mathcal{F}}(\mathbf{V}). \qquad (4.16)$$

Theorem 22. *Assume (E.1), (E.2), (E.3), (E.4), and $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$ hold. Let $\mathcal{F}$ be a parametric ensemble with continuous functions that characterizes $\mathcal{S}_{Y|\mathbf{X}}$, with $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$, and $\mathbf{V} \in S(p,q)$ defined in (1.1), with $q = p - k$. Then $\mathbf{V}_q$ in (4.16) is well defined and*

$$\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{V}_q\}^\perp. \qquad (4.17)$$

Proof of Theorem 22. Under assumptions (E.1), (E.2), and (E.3), (4.3) is well defined and continuous by using $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$ and analogous arguments to those in the proof of Theorem 8. Therefore, (4.16) 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 $\dim(\tilde{\mathcal{S}}) = \dim(\mathcal{S}_{Y|\mathbf{X}}) = k$, further let $\mathbf{Z} \in \mathbb{R}^{p\times(p-k)}$ be an orthonormal base of $\tilde{\mathcal{S}}^\perp$. Then we assume $L_\mathcal{F}(\mathbf{Z}) = \min_{V\in\mathcal{S}(p,p-k)} L_\mathcal{F}(\mathbf{V})$ and show a contradiction.

By (4.8) and (4.9) in Theorem 20, $L^*(\mathbf{V}, f_t)$, considered as a function from $\mathbb{R}^{p\times(p-k_t)}$, is minimized by an orthonormal base of $\mathcal{S}^\perp_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ with dimensions $p \times (p-k_t)$ where $k_t = \dim(\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}) \leq k$. Then from (E.1), i.e. $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} \subseteq \mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$ and the proof of Theorem 20 we see that $L^*(\mathbf{V}, f_t)$ as a function from $\mathbb{R}^{p\times(p-k)}$ is minimized by an orthonormal base $\mathbf{U} \in \mathbb{R}^{p\times(p-k)}$ of $\text{span}\{\mathbf{B}\}^\perp$.

Since $\tilde{\mathcal{S}} = \text{span}\{\mathbf{Z}\} \neq \text{span}\{\mathbf{U}\} = \mathcal{S}_{Y|\mathbf{X}}$, we can rearrange the bases $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ such that $\text{span}\{\mathbf{U}_1\} = \text{span}\{\mathbf{Z}_1\}$ and $\text{span}\{\mathbf{U}_2\} \neq \text{span}\{\mathbf{Z}_2\}$. Since $\mathcal{F}$ 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, we obtain the following contradiction

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

□

4.3 Estimation of ecve

Assume $(Y_i, \mathbf{X}_i^T)^T_{i=1,\dots,n}$ is an i.i.d. sample from model (1.13). We use $d_i(\mathbf{V}, \mathbf{s}_0)$ as in (2.21) and $w_i(\mathbf{V}, \mathbf{s}_0)$ as in (2.22) and let

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0, f) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0) f(Y_i)^l \quad \text{for} \quad l = 1, 2 \tag{4.18}$$

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

$$\tilde{L}_n(\mathbf{V}, s_0, f) = \bar{y}_2(\mathbf{V}, \mathbf{s}_0, f) - \bar{y}_1(\mathbf{V}, \mathbf{s}_0, f)^2 \tag{4.19}$$

The estimate of the objective function $L^*_\mathcal{F}(\mathbf{V}, f)$ in (4.3) is defined as

$$L^*_n(\mathbf{V}, f) = \frac{1}{n}\sum_{i=1}^n \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f), \tag{4.20}$$

where each data point $\mathbf{X}_i$ is a shifting point. For a parametric ensemble $\mathcal{F} = \{f_t : t \in \Omega_T\}$ and $(t_j)_{j=1,\dots,m_n}$ an i.i.d. sample from F_T with $\lim_{n\to\infty} m_n = \infty$, the final estimate of the objective function in (4.3) is given by

$$L_{n,\mathcal{F}}(\mathbf{V}) = \frac{1}{m_n}\sum_{j=1}^{m_n} L^*_n(\mathbf{V}, f_{t_j}) \tag{4.21}$$

Definition 23. *The* ECVE *is defined to be any basis of* $\operatorname{span}\{\hat{\mathbf{V}}_q\}^{\perp}$*, where*

$$\hat{\mathbf{V}}_q = \operatorname{argmin}_{\mathbf{V}\in S(p,q)} L_{n,\mathcal{F}}(\mathbf{V}). \tag{4.22}$$

The same algorithm as in section 2.6 is used to solve the optimization problem (4.22), which requires the gradient of (4.21). Theorem 24 provides the gradient when a Gaussian kernel is used.

Theorem 24. *The gradient of* $\tilde{L}_n(\mathbf{V}, s_0, f)$ *in* (4.19) *is given by*

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

and the gradient of $L_{n,\mathcal{F}}(\mathbf{V})$ *in* (4.21) *is*

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

The proof of Theorem 24 is analogue to the proof of Theorem 14.

For the choice of the bandwidth we use the same plug-in rule as in CVE in Section 2.4, see (2.36).

4.4 Consistency of ecve

The consistency of ECVE derives from the consistency of CVE [FB21a] 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 25, which is the main building block, to Theorem 28. The proofs are technical and lengthy, and, thus, are given in Section 4.4.1.

Theorem 25. *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 19. Then,* $L_n^*(\mathbf{V}, f)$*, defined in* (4.20)*, converges uniformly in probability to* $L^*(\mathbf{V}, f)$ *in* (4.4) *for all* $f \in \mathcal{F}$*; i.e.,*

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

Next, Theorem 26 shows that ensemble conditional variance estimator is consistent for $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ for any transformation f.

Theorem 26. *Under the same conditions as Theorem 25, the conditional variance estimator* $\operatorname{span}\{\widehat{\mathbf{B}}_{k_t}^t\}$ *estimates* $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ *consistently, for* $f_t \in \mathcal{F}$*. That is,*

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

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

$$\widehat{\mathbf{V}}^t_{k_t} = \operatorname{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L^*_n(\mathbf{V}, f_t).$$

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

A straightforward application of Theorem 26, using the identity function, obtains that $\mathcal{S}_{E(Y|\mathbf{X})}$ can be consistently estimated by CVE. We present a short juxtaposition of the assumptions in chapter 2 and 4. Assumption (A.1) corresponds to the more general assumption (E.1) in Section 4.2, i.e. (E.1) is more general, and encapsulates the model $Y = g(\mathbf{B}^T\mathbf{X}) + \tilde{\epsilon}$ with $\mathbb{E}(\tilde{\epsilon} \mid \mathbf{X}) = 0$, showing that CVE is consistent even if the mean subspace is a proper subset of the central subspace. (A.2), (A.3) correspond to assumption (E.2), (E.4). Assumption (E.3) can be removed in the context of the trivial ensemble. In total, if (A.1), (A.2), (A.3), and (A.4) hold, then (E.1), (E.2), and (E.4) hold for the univariate ensemble containing the identity function

Theorem 27. *Assume the conditions of Theorem 25 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* $(Y_i, \mathbf{X}_i)_{i=1,\dots,n}$. *Then* $L_{n,\mathcal{F}}(\mathbf{V})$, *defined in (4.21), converges uniformly in probability to* $L_{\mathcal{F}}(\mathbf{V})$ *in (4.3); i.e.,*

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

The assumption $\sup_{t\in\Omega_T} |f_t(Y)| < M < \infty$ in Theorem 27 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 25 can be dropped since obviously no truncation is needed.

The rate of convergence of m_n is not characterized in Theorem 27. In the simulation studies of Sections 4.5.2, 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 28.

Theorem 28. *Assume the conditions of Theorem 25 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* $(Y_i, \mathbf{X}_i)_{i=1,\dots,n}$. *Then, the ensemble conditional variance estimator (ECVE) is a consistent estimator for* $\mathcal{S}_{Y|\mathbf{X}}$. *That is, for any basis* $\widehat{\mathbf{B}}_{p-q,\mathcal{F}}$ *of* $\operatorname{span}\{\widehat{\mathbf{V}}_q\}^\perp$, *where* $\widehat{\mathbf{V}}_q$ *is defined in (4.22) with* $q = p - k$ *and* $k = \dim(\mathcal{S}_{Y|\mathbf{X}})$,

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

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

4.4.1 Proofs

Road-map for the proof of consistency.

1. Theorem 19 shows that we can decompose the transformed response $f(Y) = g(\widetilde{\mathbf{B}}^T\mathbf{X}) + \tilde{\epsilon}$ such that it fits into the framework of CVE in chapter 2.

2. Theorem 20 shows that CVE can identify $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ on the population level, i.e. $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \text{span}\{\text{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L^*(\mathbf{V}, f_t)\}^\perp$.

3. Theorem 22 states that the target function $L_{\mathcal{F}}(\mathbf{V})$ of ECVE identifies $\mathcal{S}_{Y|\mathbf{X}}$ on the population level if $\mathcal{F}$ characterises $\mathcal{S}_{Y|\mathbf{X}}$, i.e. $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\text{argmin}_{\mathbf{V}\in\mathcal{S}(p,q)} L_{\mathcal{F}}(\mathbf{V})\}^\perp$

4. Suppose f is a fixed arbitrary element of $\mathcal{F}$ and by Theorem 19

$$\tilde{Y}_i = f(Y_i) = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i) + \tilde{\epsilon}_i \tag{4.23}$$

   with $\text{span}\{\widetilde{\mathbf{B}}\} = \mathcal{S}_{\mathbb{E}(\tilde{Y}|\mathbf{X})} = \mathcal{S}_{\mathbb{E}(f(Y)|\mathbf{X})}$. Condition (E.4) yields that g is twice continuously differentiable, and $\mathbb{E}(|\tilde{Y}|^8) < \infty$.

5. Since f is fixed, we suppress the dependence on the transformation f in the notation of $t^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = t^{(l)}(\mathbf{V}, \mathbf{s}_0)$ given in (2.6), $\tilde{h}(\mathbf{V}, \mathbf{s}_0, f) = \tilde{h}(\mathbf{V}, \mathbf{s}_0)$ given in (4.7), $\tilde{L}(\mathbf{V}, \mathbf{s}_0, f) = \tilde{L}(\mathbf{V}, \mathbf{s}_0)$ given in (4.2), and their corresponding sample counterparts for convenience till the proof of Theorem 25[2]. We set

$$t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) = t_n^{(l)}(\mathbf{V}, \mathbf{s}_0, f) = \frac{1}{nh_n^{(p-q)/2}} \sum_{i=1}^n K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) \tilde{Y}_i^l, \tag{4.24}$$

   which is the sample version of (2.6) for $l = 0, 1, 2$. Eqn. (4.18) can be expressed as

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \sum_{i=1}^n w_i(\mathbf{V}, \mathbf{s}_0)\tilde{Y}_i^l = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)}. \tag{4.25}$$

6. The strategy is to show uniform convergence in probability of $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$ to its population counterpart $t^{(l)}(\mathbf{V}, \mathbf{s}_0)$[3] in lemma 34, i.e. $\sup_{\mathbf{V}\in\mathcal{S}(p,q), \mathbf{s}_0\in\text{supp}(f_{\mathbf{X}})} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t^{(l)}(\mathbf{V}, \mathbf{s}_0)| \to 0$ in probability.

   This will be done via the inequality

$$\begin{aligned}\sup_{\mathbf{V}\in\mathcal{S}(p,q), \mathbf{s}_0\in\text{supp}(f_{\mathbf{X}})} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t^{(l)}(\mathbf{V}, \mathbf{s}_0)| &\le \sup_{\mathbf{V}, \mathbf{s}_0} |t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0))| \\ &+ \sup_{\mathbf{V}, \mathbf{s}_0} |\mathbb{E}(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)) - t^{(l)}(\mathbf{V}, \mathbf{s}_0)|.\end{aligned}$$

[2] To be precise, only in the proof of Theorem 27 the dependence on the transformation f is relevant since before we just have an arbitrary but fixed function f.

[3] For $l = 2$ the population counterpart is $t^{(2)}(\mathbf{V}, \mathbf{s}_0, f) + \tilde{h}(\mathbf{V}, \mathbf{s}_0, f)t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ due to $\tilde{Y}_i^2 = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)^2 + 2g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)\tilde{\epsilon}_i + \tilde{\epsilon}_i{}^2$, i.e. the cross term vanishes but the $\tilde{\epsilon}_i{}^2$ term yields the additional $\tilde{h}(\mathbf{V}, \mathbf{s}_0, f)t^{(0)}(\mathbf{V}, \mathbf{s}_0)$.

The first term on the right hand side above deals with the random fluctuations of $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$ about its expectation. Lemma 32 shows that this converges to 0 using the Bernstein inequality (4.30).

The second term on the right hand side is a bias, i.e. the difference between expectation and population quantity. This term will be handled in lemma 33 using traditional techniques from kernel density estimation.

7. Lemmas 29, 30, and 31 are auxiliary lemmas.

8. Lemma 35 is an auxiliary lemma used in the proof of Theorem 36.

9. Theorem 36 yields that

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

uniformly in probability[4] by utilizing lemma 34, the continuity of $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$ on $\mathcal{S}(p, q) \times \{\mathbf{x} : f_{\mathbf{X}}(\mathbf{x}) > 0\}$ due to the kernel being continuous by (K.1), and lemma 35.

Furthermore, it yields that (4.19) converges to (4.2), i.e. $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0) \to \tilde{L}(\mathbf{V}, \mathbf{s}_0)$, uniformly in probability.

10. Theorem 37 is Theorem 2 of [Jen69] or [MMW+63, p. 40], which will be used in the proof of Theorems 25 and 27.

11. Theorem 25 shows that the target function $L_n^*(\mathbf{V}, f)$ defined in (4.4) of CVE converges uniformly in $\mathbf{V} \in \mathcal{S}(p, q)$ to $L^*(\mathbf{V}, f)$, defined in (4.4), in probability

12. Theorem 26 establish that CVE estimates $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})}$ consistently by utilizing the uniform convergence in probability of its target function $L_n^*(\mathbf{V}, f)$ defined in (4.4).

13. Theorem 27 establishes that the target function $L_{n,\mathcal{F}}(\mathbf{V})$ defined in (4.21) of ECVE converges uniformly in $\mathbf{V} \in \mathcal{S}(p, q)$ to $L_{\mathcal{F}}(\mathbf{V})$, defined in (4.3), in probability.

    The proof uses the Markov inequality, Fatou's lemma by means of the assumption $\sup_{t \in \Omega_T} |f_t(Y)| < M < \infty$, and Theorem 26.

14. Theorem 28 shows that ECVE estimates $\mathcal{S}_{Y|\mathbf{X}}$ consistently by utilizing the uniform convergence in probability of its target function shown in Theorem 27.

First we introduce notation and auxiliary lemmas for the proof of Theorem 25. We suppose all assumptions of Theorem 25 hold. We generically use the letter "C" to denote constants.

Lemma 29. *Assume the conditions of Theorem 25 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,*

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

[4] For $l = 2$ we again get the extra term, i.e. $\bar{y}_2(\mathbf{V}, \mathbf{s}_0) \to \mu_2(\mathbf{V}, \mathbf{s}_0) + \tilde{h}(\mathbf{V}, \mathbf{s}_0)$ as explained in point 6 of the road-map.

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

Proof of Lemma 29. By (1.3), $\|\mathbf{P}_{\mathbf{U}}(\mathbf{x} - \mathbf{s}_0)\|^2 = \|\mathbf{U}\mathbf{r}_2\|^2 = \|\mathbf{r}_2\|^2$. Further

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

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

Lemma 30. *Assume the conditions of Theorem 25 hold. Then, there exists a constant $C > 0$, such that*

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

for $n > n^\star$ and $t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)$, $l = 0, 1, 2$, in (4.24).

Proof of Lemma 30. Since a continuous function attains a finite maximum over a compact set, $\sup_{\mathbf{x} \in \mathrm{supp}(f_{\mathbf{X}})} |g(\widetilde{\mathbf{B}}^T \mathbf{x})| < \infty$. Therefore,

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

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

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left(nh_n^{(p-q)/2} t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right) &= n\mathbb{V}\mathrm{ar}\left(\tilde{Y}^l K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \leq n\mathbb{E}\left(\tilde{Y}^{2l} K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \\
&= n\mathbb{E}\left(|\tilde{Y}|^{2l} K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \leq n \sum_{u=0}^{2l} \binom{2l}{u} C^u \mathbb{E}\left(|\tilde{\epsilon}_i|^{2l-u} K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \\
&= n \sum_{u=0}^{2l} \binom{2l}{u} C^u \mathbb{E}\left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i) K^2\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right)\right) \qquad (4.26)
\end{aligned}$$

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 $\mathrm{supp}(f_{\mathbf{X}})$. Using Lemma 29 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

$$\begin{aligned}\mathbb{E}\left(\mathbb{E}(|\tilde{\epsilon}_i|^{2l-u} \mid \mathbf{X}_i)K^2\left(\frac{d_i(\mathbf{V},\mathbf{s}_0)}{h_n}\right)\right) &= h_n^{(p-q)/2}\mathbb{E}(Z_n(\mathbf{V},\mathbf{s}_0)) \\ &= h_n^{(p-q)/2}\int_{\text{supp}(f_{\mathbf{X}})\cap\mathbb{R}^{p-q}} K^2(\|\mathbf{r}_2\|^2)\times \\ \int_{\text{supp}(f_{\mathbf{X}})\cap\mathbb{R}^q} g_{2l-u}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+h_n^{1/2}\mathbf{U}\mathbf{r}_2)f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1+h_n^{1/2}\mathbf{U}\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 &\qquad (4.27)\\ &\leq h_n^{(p-q)/2}C\end{aligned}$$

since all integrands in (4.27) 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 (4.27) into (4.26) yields

$$\mathbb{V}\text{ar}\left(nh_n^{(p-q)/2}t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right) \leq nh_n^{(p-q)/2}\underbrace{\sum_{u=0}^{2l}\binom{2l}{u}C^u C}_{=C} = nh_n^{(p-q)/2}C \qquad (4.28)$$

□

In Lemma 31 we show that $d_i(\mathbf{V},\mathbf{s}_0)$ in (2.21) is Lipschitz in its inputs under assumption (E.2).

Lemma 31. *Under assumption (E.2) there exists a constant $0 < C_2 < \infty$ such that for all $\delta > 0$ and $\mathbf{V},\mathbf{V}_j \in \mathcal{S}(p,q)$ with $\|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| < \delta$ and for all $\mathbf{s}_0,\mathbf{s}_j \in supp(f_{\mathbf{X}}) \subset \mathbb{R}^p$ with $\|\mathbf{s}_0 - \mathbf{s}_j\| < \delta$,*

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

for $d_i(\mathbf{V},\mathbf{s}_0)$ given by (2.21)

Proof of Lemma 31.

$$\begin{aligned}|d_i(\mathbf{V},\mathbf{s}_0) - d_i(\mathbf{V}_j,\mathbf{s}_j)| &\leq \left|\|\mathbf{X}_i-\mathbf{s}_0\|^2 - \|\mathbf{X}_i-\mathbf{s}_j\|^2\right| + \\ \left|\langle \mathbf{X}_i-\mathbf{s}_0, \mathbf{P}_{\mathbf{V}}(\mathbf{X}_i-\mathbf{s}_0)\rangle - \langle \mathbf{X}_i-\mathbf{s}_j, \mathbf{P}_{\mathbf{V}_j}(\mathbf{X}_i-\mathbf{s}_j)\rangle\right| &= I_1 + I_2 \qquad (4.29)\end{aligned}$$

where $\langle\cdot,\cdot\rangle$ is the scalar product in $\mathbb{R}^p$. We bound the first term on the right hand side of (4.29) as follows using $\|\mathbf{X}_i\| \leq \sup_{z\in\text{supp}(f_{\mathbf{X}})}\|z\| = C_1 < \infty$ with probability 1 by (E.2).

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

by Cauchy-Schwartz and the reverse triangular inequality for which $\left|\|\mathbf{s}_0\|^2 - \|\mathbf{s}_j\|^2\right| = |\|\mathbf{s}_0\| - \|\mathbf{s}_j\|| \left(\|\mathbf{s}_0\| + \|\mathbf{s}_j\|\right) \leq \|\mathbf{s}_0-\mathbf{s}_j\|2C_1$. The second term in (4.29) satisfies

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

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

To show Theorem 25 and Lemma 32, 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,

$$P(|S_n - E_n| > t) < 2\exp\left(-\frac{t^2/2}{V_n + bt/3}\right) \tag{4.30}$$

Assumption (K.2) yields

$$|K(u) - K(u')| \leq K^*(u')\delta \tag{4.31}$$

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 32 and 33 we show that (4.24) converges uniformly in probability to (2.6) by showing that the variance and bias terms vanish uniformly in probability, respectively.

Lemma 32. *Under the assumptions of Theorem 25,*

$$\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| = O_P(a_n), \; l = 0, 1, 2 \tag{4.32}$$

Proof of Lemma 32. 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 (4.30). If the function f in (4.23) 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

$$t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) = (1/nh_n^{(p-q)/2}) \sum_i K(\|\mathbf{P}_{\mathbf{U}}(\mathbf{X}_i - \mathbf{s}_0)\|^2 / h_n) \tilde{Y}_i^l 1_{\{|\tilde{Y}_i|^l \leq \tau_n\}} \tag{4.33}$$

be the truncated version of (4.24) 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 (4.24). 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

$$\begin{aligned}\sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| &\leq M_1(\tilde{R}_n^{(l)} + \mathbb{E}\tilde{R}_n^{(l)}) \\ &+ \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_{n,\text{trc}}^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right|\end{aligned} \tag{4.34}$$

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

$$\begin{aligned}\mathbb{E}\tilde{R}_n^{(l)} &= \frac{1}{h_n^{(p-q)/2}} \mathbb{E}\left(|\tilde{Y}_i|^l 1_{\{|\tilde{Y}_i|^l > \tau_n\}}\right) \leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})} \sqrt{\mathbb{P}(|\tilde{Y}_i|^l > \tau_n)} \\ &\leq \frac{1}{h_n^{(p-q)/2}} \sqrt{\mathbb{E}(|\tilde{Y}_i|^{2l})} \left(\frac{\mathbb{E}(|\tilde{Y}_i|^{4l})}{a_n^{-4}}\right)^{1/2} = o(a_n),\end{aligned} \tag{4.35}$$

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 (4.34) converge to 0 in probability with rate a_n by (4.35) 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 $A_j = \{\mathbf{V} : \|\mathbf{P}_{\mathbf{V}} - \mathbf{P}_{\mathbf{V}_j}\| \leq \delta_n\} \times \{\mathbf{s} : \|\mathbf{s} - \mathbf{s}_j\| \leq \delta_n\}$ 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 31 there exists $0 < C_2 < \infty$, such that

$$|d_i(\mathbf{V}, \mathbf{s}_0) - d_i(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2\delta_n \tag{4.36}$$

for d_i in (2.21). Under (K.2), which implies (4.31), inequality (4.36) yields

$$\left|K\left(\frac{d_i(\mathbf{V}, \mathbf{s}_0)}{h_n}\right) - K\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right)\right| \leq K^*\left(\frac{d_i(\mathbf{V}_j, \mathbf{s}_j)}{h_n}\right) C_2 a_n \tag{4.37}$$

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 (4.37) yields

$$|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)| \leq C_2 a_n r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) \tag{4.38}$$

Since K^* fulfills (K.1) except for continuity, an analogous argument as in the proof of Lemma 29 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, (4.38) and integrability of r_n^l, we obtain

$$\begin{aligned}
&\left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \leq \left|t_n^{(l)}(\mathbf{V}, \mathbf{s}_0) - t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right| + \left|\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)\right)\right| \\
&+ \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \leq C_2 a_n \left(|r_n| + |\mathbb{E}\left(r_n\right)|\right) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq C_2 a_n(|r_n - \mathbb{E}(r_n)| + 2|\mathbb{E}(r_n)|) + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right| \\
&\leq 2C_3 a_n + |r_n - \mathbb{E}(r_n)| + \left|t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j) - \mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j)\right)\right|
\end{aligned} \tag{4.39}$$

for any constant $C_3 > C_2\mathbb{E}(r_n^{(l)}(\mathbf{V}_j, \mathbf{s}_j))$ and n such that $C_2 a_n \leq 1$, since $a_n^2 = o(1)$, which in turn yields that there exists $0 < C_3 < \infty$ such that (4.39) holds.

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

and continuous function f,

$$\begin{aligned}
&\mathbb{P}(\sup_{\mathbf{V}\times\mathbf{s}_0\in A}|t_n^{(l)}(\mathbf{V},\mathbf{s}_0)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)|>3C_3a_n)\\
&\leq\sum_{j=1}^{N}\mathbb{P}(\sup_{\mathbf{V}\times\mathbf{s}_0\in A_j}|t_n^{(l)}(\mathbf{V},\mathbf{s}_0)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)|>3C_3a_n)\\
&\leq N\max_{1\leq j\leq N}\mathbb{P}(\sup_{\mathbf{V}\times\mathbf{s}_0\in A_j}|t_n^{(l)}(\mathbf{V},\mathbf{s}_0)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)|>3C_3a_n) \qquad (4.40)\\
&\leq N\left(\max_{1\leq j\leq N}\mathbb{P}(|t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)\right)|>C_3a_n)+\max_{1\leq j\leq N}\mathbb{P}(|r_n-\mathbb{E}(r_n)|>C_3a_n)\right)\leq\\
&C\,\delta^{-(d+p)}\left(\max_{1\leq j\leq N}\mathbb{P}(|t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)\right)|>C_3a_n)+\max_{1\leq j\leq N}\mathbb{P}(|r_n-\mathbb{E}(r_n)|>C_3a_n)\right)
\end{aligned}$$

by the subadditivity of probability for the first inequality and (4.39) for the third inequality above, where the last inequality is due to $N\leq C\,\delta_n^{-d}\delta_n^{-p}$ for a cover of A.

Finally, we bound the first and second term in the last line of (4.40) by the Bernstein inequality (4.30). For the first term in the last line of (4.40), let $Z_i=Y_i^lK(d_i(\mathbf{V}_j,\mathbf{s}_j)/h_n)$ and $S_n=\sum_i Z_i=nh_n^{(p-q)/2}t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)$. Then, Z_i are independent with $|Z_i|\leq b=M_1\tau_n=M_1/a_n$ by (K.1) and the truncation step (i). For $V_n=\mathbb{V}\mathrm{ar}(S_n)$, Lemma 30 yields $nh_n^{(p-q)/2}C\geq V_n$ with $C>0$, and set $t=C_3a_nnh_n^{(p-q)/2}$. The Bernstein inequality (4.30) yields

$$\begin{aligned}
&\mathbb{P}\left(\left|t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)\right)\right|>C_3a_n\right)<2\exp\left(\frac{-t^2/2}{V_n+bt/3}\right)\leq\\
&2\exp\left(-\frac{(1/2)C_3^2a_n^2n^2h_n^{(p-q)}}{nh_n^{(p-q)/2}C+(1/3)M_1\tau_nC_3a_nnh_n^{(p-q)/2}}\right)\leq 2\exp\left(-\frac{(1/2)C_3\log(n)}{C/C_3+M_1/3}\right)=2n^{-\gamma(C_3)}
\end{aligned}$$

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 (4.40), set $Z_i=Y_i^lK^*(d_i(\mathbf{V}_j,\mathbf{s}_j)/h_n)$ in (4.30) and proceed similarly to obtain

$$\mathbb{P}\left(\left|r_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)-\mathbb{E}\left(r_n^{(l)}(\mathbf{V}_j,\mathbf{s}_j)\right)\right|>C_3a_n\right)<2n^{-\frac{(1/2)C_3}{C/C_3+(1/3)M_2}}=2n^{-\gamma(C_3)}$$

By (H.1), $h_n^{(p-q)/4}\leq 1$ for n large and (H.2) implies $1/(nh_n^{(p-q)/2})\leq 1$ for n large, therefore $h_n^{-1}\leq n^{2/(p-q)}\leq n^2$ since $p-q\geq 1$. Then, $\delta_n^{-1}=(a_nh_n)^{-1}\leq n^{1/2}h_n^{-1}h_n^{(p-q)/4}\leq n^{5/2}$. Therefore, (4.40) is smaller than $4\,C\,\delta_n^{-(d+p)}n^{-\gamma(C_3)}\leq 4Cn^{5(d+p)/2-\gamma(C_3)}$. For C_3 large enough, we have $5(d+p)/2-\gamma(C_3)<0$ and $n^{5(d+p)/2-\gamma(C_3)}\to 0$. This completes the proof. □

If we assume $|\tilde{Y}_i|<M_2<\infty$ almost surely, the requirement $a_n/h_n^{(p-q)/2}=O(1)$ for the bandwidth can be dropped and the truncation step of the proof of Lemma 32 is no longer necessary.

Lemma 33. *Assume the conditions of Theorem 25, and* $\int_{\mathbb{R}^{p-q}} K(\|\mathbf{r}_2\|^2)d\mathbf{r}_2 = 1$ *hold, then*

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A}\left|t^{(l)}(\mathbf{V},\mathbf{s}_0)+1_{\{l=2\}}\tilde{h}(\mathbf{V},\mathbf{s}_0)t^{(0)}(\mathbf{V},\mathbf{s}_0)-\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right)\right| = O(h_n), \quad l=0,1,2 \tag{4.41}$$

where $t^{(l)}(\mathbf{V},\mathbf{s}_0)$ *and* $\tilde{h}(\mathbf{V},\mathbf{s}_0)$ *are defined in Theorem 20.*

Proof of Lemma 33. 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 (1.3).

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

Then

$$\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right) = \int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\tilde{g}(\mathbf{r}_1,{h_n}^{1/2}\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2 \tag{4.42}$$

holds by Lemma 29 for $l=0,1$. Plugging in (4.42) the second order Taylor expansion for some ξ in the neighborhood of 0, $\tilde{g}(\mathbf{r}_1,{h_n}^{1/2}\mathbf{r}_2) = \tilde{g}(\mathbf{r}_1,0)+{h_n}^{1/2}\nabla_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,0)^T\mathbf{r}_2+h_n\mathbf{r}_2^T\nabla^2_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,\xi)\mathbf{r}_2$, yields

$$\begin{aligned}
\mathbb{E}\left(t_n^{(l)}(\mathbf{V},\mathbf{s}_0)\right) &= \int_{\mathbb{R}^q}\tilde{g}(\mathbf{r}_1,0)d\mathbf{r}_1+\sqrt{h_n}\left(\int_{\mathbb{R}^q}\nabla_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,0)d\mathbf{r}_1\right)^T\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\mathbf{r}_2d\mathbf{r}_2+\\
&\quad h_n\frac{1}{2}\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\mathbf{r}_2^T\nabla^2_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,\xi)\mathbf{r}_2d\mathbf{r}_1d\mathbf{r}_2 = t^{(l)}(\mathbf{V},\mathbf{s}_0)+h_n\frac{1}{2}R(\mathbf{V},\mathbf{s}_0)
\end{aligned}$$

since $\int_{\mathbb{R}^q}\tilde{g}(\mathbf{r}_1,0)d\mathbf{r}_1 = t^{(l)}(\mathbf{V},\mathbf{s}_0)$ and $\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\mathbf{r}_2d\mathbf{r}_2 = 0\in\mathbb{R}^{p-q}$ due to $K(\|\cdot\|^2)$ being even. Let $R(\mathbf{V},\mathbf{s}_0) = \int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\int_{\mathbb{R}^p}\mathbf{r}_2^T\nabla^2_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,\xi)\mathbf{r}_2d\mathbf{r}_1d\mathbf{r}_2$. By (E.4) and (E.2), $|\mathbf{r}_2^T\nabla^2_{\mathbf{r}_2}\tilde{g}(\mathbf{r}_1,\xi)\mathbf{r}_2|\le C\|\mathbf{r}_2\|^2$ for $C=\sup_{\mathbf{x},\mathbf{y}}\|\nabla^2_{\mathbf{r}_2}\tilde{g}(\mathbf{x},\mathbf{y})\|<\infty$, since a continuous function over a compact set is bounded. Then, $R(\mathbf{V},\mathbf{s}_0)\le CC_4\int_{\mathbb{R}^{p-q}}K(\|\mathbf{r}_2\|^2)\|\mathbf{r}_2\|^2d\mathbf{r}_2<\infty$ for some $C_4>0$, since the integral over $\mathbf{r}_1$ is over a compact set by (E.2).

For $l=2$, $\tilde{Y}_i^2 = g_i^2+2g_i\tilde{\epsilon}_i+\tilde{\epsilon}_i^2$ with $g_i = g(\widetilde{\mathbf{B}}^T\mathbf{X}_i)$, and this case can be handled completely analogue as for $l=0,1$. The term $\tilde{\epsilon}_i^2$ yields the extra term

$$1_{\{l=2\}}\tilde{h}(\mathbf{V},\mathbf{s}_0)t^{(0)}(\mathbf{V},\mathbf{s}_0) = 1_{\{l=2\}}\int_{\mathrm{supp}(f_{\mathbf{X}})\cap\mathbb{R}^q}h(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)f_{\mathbf{X}}(\mathbf{s}_0+\mathbf{V}\mathbf{r}_1)d\mathbf{r}_1$$

with $h(\mathbf{x}) = \mathbb{E}(\tilde{\epsilon}^2 \mid \mathbf{X} = \mathbf{x})$ given in Theorem 20. □

Lemma 34 follows directly from Lemmas 32 and 33 and the triangle inequality.

Lemma 34. *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* $l = 0, 1, 2$

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

Next we present lemma 35 that will be used in the proof of Theorem 36.

Lemma 35. *Let* $A \subseteq \mathcal{X}$ *be a bounded set and* $f_n(\mathbf{x})$, *from an Euclideaqn space* $\mathcal{X}$ *to* $\mathbb{R}$, *be a sequence of continuous functions. Further let* $A_n \uparrow A$ *be an increasing sequence of sets such that the Hausdorff distance* $|A_n - A| \to 0$ *for* $n \to \infty$. *Then*

$$\lim_{n\to\infty} \sup_{\mathbf{x}\in A_n} f_n(\mathbf{x}) = \lim_{n\to\infty} \sup_{\mathbf{x}\in A} f_n(\mathbf{x})$$

Proof of lemma 35. Since $A_n \subseteq A$, we have $\sup_{\mathbf{x}\in A_n} f_n(\mathbf{x}) \leq \sup_{\mathbf{x}\in A} f_n(\mathbf{x})$. For the other inequality, let $\nu > 0$ and denote $s_n = \sup_{\mathbf{x}\in A_n} f_n(\mathbf{x})$ and $s_n^* = \sup_{\mathbf{x}\in A} f_n(\mathbf{x})$. Note that by the continuity of f_n and definition of the supremum there exists a sequence $x_{m,n}^* \in A$ such that $f_n(x_{m,n}^*) \uparrow s_n^*$ for $m \to \infty$. Especially there is a integer M such that $|f_n(x_{m,n}^*) - s_n^*| < \nu/2$ for $m > M$. Fix some $m > M$ and write $x_n^* = x_{m,n}^*$.

Moreover, by the continuity of f_n at x_n^*, there is a $\delta > 0$ such that $|f_n(x) - f_n(x_n^*)| < \nu/2$ for all x with $|x - x_n^*| < \delta$. Then choose n so large that $|A_n - A| < \delta$, therefore there exists $x_n \in A_n$ such that $|x_n - x_n^*| < \delta$ and by the continuity of f_n it holds $s_n^* - f_n(x_n) = |f_n(x_n^*) - f_n(x_n)| < \nu$. Rearranging yields $\sup_{\mathbf{x}\in A} f_n(\mathbf{x}) - \nu = s_n^* - \nu < f_n(x_n) \leq \sup_{\mathbf{x}\in A_n} f_n(\mathbf{x})$ and since ν was arbitrary the other inequality follows, completing the proof. □

Theorem 36. *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*

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

and

$$\sup_{\mathbf{V}\times\mathbf{s}_0\in A} \left| \tilde{L}_n(\mathbf{V}, \mathbf{s}_0) - \tilde{L}(\mathbf{V}, \mathbf{s}_0) \right| = o_P(1) \tag{4.43}$$

where $\bar{y}_l(\mathbf{V}, \mathbf{s}_0)$, $\mu_l(\mathbf{V}, \mathbf{s}_0)$, $\tilde{L}_n(\mathbf{V}, \mathbf{s}_0)$ *and* $\tilde{L}(\mathbf{V}, \mathbf{s}_0)$ *are defined in* (4.18), (2.5), (4.19) *and* (4.6), *respectively.*

Proof of Theorem 36. 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 (2.6), 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. Note that $t^{(0)}(\mathbf{V}, \mathbf{s}_0) = \int_{\text{supp}(f_{\mathbf{X}})\cap\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1)d\mathbf{r}_1 > 0$ if $f_{\mathbf{X}}(\mathbf{s}_0) > 0$ by the convex support of the continuous density (see proof of Theorem 8). $t^{(0)}(\mathbf{V}, \mathbf{s}_0)$ can only be 0 if $f_{\mathbf{X}}(\mathbf{s}_0) = 0$, i.e. $\mathbf{s}_0 \in \partial\text{supp}(f_{\mathbf{X}})$

and the directions $\mathbf{V}$ are tangential to the boundary (i.e. think of the support being a circle and $\mathbf{s}_0$ being on the boundary and $\mathbf{V}$ tangential, then $t^{(0)}(\mathbf{V}, \mathbf{s}_0) = 0$ since we integrate only over one boundary point of the support of $f_{\mathbf{X}}$), therefore $\delta_n \to 0$ as $n \to \infty$. Then,

$$\bar{y}_l(\mathbf{V}, \mathbf{s}_0) = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)} = \frac{t_n^{(l)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)}{t_n^{(0)}(\mathbf{V}, \mathbf{s}_0)/t^{(0)}(\mathbf{V}, \mathbf{s}_0)} \tag{4.44}$$

We consider the numerator and enumerator of (4.44) separately. By Lemma 34

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

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

Then $A_n \uparrow A = \mathcal{S}(p, q) \times \text{supp}(f_{\mathbf{X}})$ and Corollary 5.3 of [Tuz20], yield $A_n \to A$ with respect to the Hausdorff distance, by (K.1) $t_n^{(l)}$ is continuous, by the proof of Theorem 8 $t^{(l)}$ is continuous, and applying Lemma 35, yields

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

Substituting in (4.44), we obtain uniformly in $\mathbf{V}, \mathbf{s}_0$

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

The case $l = 2$, can be handled analogously and by Lemma 34 yields the extra term $\bar{y}_2(\mathbf{V}, \mathbf{s}_0) = t_n^{(2)}(\mathbf{V}, \mathbf{s}_0)/t_n^{(0)}(\mathbf{V}, \mathbf{s}_0) \to \left(t^{(2)}(\mathbf{V}, \mathbf{s}_0) + \tilde{h}(\mathbf{V}, \mathbf{s}_0)t^{(0)}(\mathbf{V}, \mathbf{s}_0)\right)/t^{(0)}(\mathbf{V}, \mathbf{s}_0) = \mu_2(\mathbf{V}, \mathbf{s}_0) + \tilde{h}(\mathbf{V}, \mathbf{s}_0)$. This shows the first statement of Theorem 36, i.e. $\bar{y}_l(\mathbf{V}, \mathbf{s}_0)$ converges uniformly in probability to its population counterpart.

Further, to obtain the second statement (4.43), note that

$$\sup_{\mathbf{V},\mathbf{s}_0} |\bar{y}_1(\mathbf{V}, \mathbf{s}_0)^2 - \mu_1(\mathbf{V}, \mathbf{s}_0)^2| \leq \sup_{\mathbf{V},\mathbf{s}_0} |\bar{y}_1(\mathbf{V}, \mathbf{s}_0) - \mu_1(\mathbf{V}, \mathbf{s}_0)| \sup_{\mathbf{V},\mathbf{s}_0} |\bar{y}_1(\mathbf{V}, \mathbf{s}_0) + \mu_1(\mathbf{V}, \mathbf{s}_0)|$$

where the first term on the right hand side goes to 0 in probability by the first statment of the theorem. The second term is bounded in probability since

$$\mu_1(\mathbf{V}, \mathbf{s}_0) = \frac{\int g(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}}{\int f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}} \leq \sup_{\mathbf{x}\in\text{supp}(f_{\mathbf{X}})} |g(\mathbf{x})| = C < \infty$$

with a finite constant C by continuity of g (E.4) and compact support of the density (E.2). By the first statement $\bar{y}_1(\mathbf{V}, \mathbf{s}_0) \to \mu_1(\mathbf{V}, \mathbf{s}_0)$ uniformly in probability, therefore $\sup_{\mathbf{V},\mathbf{s}_0} |\bar{y}_1(\mathbf{V}, \mathbf{s}_0)| \leq C + \delta$ with high probability if n is sufficiently large.

Then (4.43) follows from (4.6). □

Next we present Theorem 37 given in Theorem 2 of [Jen69] in [MMW+63, p. 40], which will be used in the proof of Theorems 25 and 27.

Theorem 37. *Let f be a function on $\mathcal{X} \times \Theta$ where $\mathcal{X}$ is a Euclidean space and Θ is a compact subset of a Euclidean space. Let $f(\mathbf{x}, \boldsymbol{\theta})$ be a continuous function of $\boldsymbol{\theta}$ for each $\mathbf{x}$ and a measurable function of $\mathbf{x}$ for each $\boldsymbol{\theta}$. Assume also that $f(\mathbf{x}, \boldsymbol{\theta}) < h(\mathbf{x})$ for all $\mathbf{x}$ and $\boldsymbol{\theta}$, where h is integrable with respect to a probability distribution function F on $\mathcal{X}$. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be an independent and identical distributed sample of F, then*

$$\frac{1}{n}\sum_{j=1}^{n} f(\mathbf{X}_i, \boldsymbol{\theta}) \longrightarrow \mathbb{E}\left(f(\mathbf{X}_1, \boldsymbol{\theta})\right) \quad \textit{uniformly over } \boldsymbol{\theta} \in \Theta \textit{ almost surely as } n \to \infty$$

Proof of Theorem 25. By (4.20) and (4.3),

$$\begin{aligned} |L_n^*(\mathbf{V}, f) - L^*(\mathbf{V}, f)| \leq & \left| \frac{1}{n} \sum_i \left(\tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}(\mathbf{V}, \mathbf{X}_i, f) \right) \right| \\ & + \left| \frac{1}{n} \sum_i \left(\tilde{L}(\mathbf{V}, \mathbf{X}_i, f) - \mathbb{E}(\tilde{L}(\mathbf{V}, \mathbf{X}, f)) \right) \right| \end{aligned} \tag{4.45}$$

By Theorem 36,

$$\left| \frac{1}{n} \sum_i \tilde{L}_n(\mathbf{V}, \mathbf{X}_i, f) - \tilde{L}(\mathbf{V}, \mathbf{X}_i, f) \right| \leq \sup_{\mathbf{V} \times \mathbf{s}_0 \in A} \left| \tilde{L}_n(\mathbf{V}, \mathbf{s}_0, f) - \tilde{L}(\mathbf{V}, \mathbf{s}_0, f) \right| = o_P(1) \tag{4.46}$$

For the second term in (4.45) we apply Theorem 37 with $f(\mathbf{X}_i, \mathbf{V}) = \tilde{L}(\mathbf{V}, \mathbf{X}_i, f)$ where $\mathbf{V} \in \mathcal{S}(p, q) \subseteq \mathbb{R}^{pq}$ is a compact subset of Euclidean space. By Theorems 19 and 20 $\mu_2(\mathbf{V}, \mathbf{s}_0, f) - \mu_1(\mathbf{V}, \mathbf{s}_0, f)^2 + \mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\})$, and by (E.2) and (E.4) we have $C = \sup_{\mathbf{x} \in \mathrm{supp}(f_{\mathbf{X}})} |g(\mathbf{x})|^l < \infty$ since a continuous function attains a finite maximum over a compact set. Therefore,

$$\mu_l(\mathbf{V}, \mathbf{s}_0, f) = \frac{\int g(\mathbf{s}_0 + \mathbf{V}\mathbf{r})^l f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}}{\int f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}} \leq C < \infty$$

, and analogue for

$$\mathbb{V}\mathrm{ar}(\tilde{\epsilon} \mid \mathbf{X} \in \mathbf{s}_0 + \mathrm{span}\{\mathbf{V}\}) = \frac{\int_{\mathrm{supp}(f_{\mathbf{X}}) \cap \mathbb{R}^q} h(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}_1) d\mathbf{r}_1}{\int_{\mathbb{R}^q} f_{\mathbf{X}}(\mathbf{s}_0 + \mathbf{V}\mathbf{r}) d\mathbf{r}} \leq \sup_{\mathbf{x} \in \mathrm{supp}(f_{\mathbf{X}})} |h(\mathbf{x})| < \infty$$

with $h(\mathbf{x})$ continuous by the assumptions of Theorem 25. Therefore $\tilde{L}(\mathbf{V}, \mathbf{X}_i, f)$ is upper bounded by a constant which is integrable and therefore Theorem 37 yields that the second term in (4.45) converges uniformly to 0 almost surely.

In total, $\sup_{\mathbf{V} \in \mathcal{S}(p,q)} |L_n^*(\mathbf{V}, f) - L^*(\mathbf{V}, f)| \leq o_P(1)$ which implies Theorem 25. □

Proof of Theorem 26. 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., $\mathbf{P}_{\operatorname{span}\{\widehat{\mathbf{B}}_{k_t}^t\}^\perp} = \operatorname{argmin} L_n^*(\mathbf{P}_\mathbf{V}, f) \to \mathbf{P}_{\operatorname{span}\{\mathbf{B}\}^\perp} = \operatorname{argmin} L^*(\mathbf{P}_\mathbf{V}, f_t)$ 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 $\mathcal{S}^\perp_{\mathbb{E}(f_t(Y)|\mathbf{X})}$.

Since $L_n^*(\mathbf{V}, f_t)$ depends on $\mathbf{V}$ only through $\mathbf{P}_\mathbf{V} = \mathbf{V}\mathbf{V}^T$, $L_n^*(\mathbf{V}, f_t)$ can be considered as functions on the Grassmann manifold, which is compact, and the same holds true for $L^*(\mathbf{V}, f_t)$ by (2.19). 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 25 obtains the uniform convergence and Theorem 20 that the minimizer is unique when $L(\mathbf{V})$ is minimized over the Grassmann manifold $G(p, q)$, since $\mathcal{S}_{\mathbb{E}(f_t(Y)|\mathbf{X})} = \operatorname{span}\{\widetilde{\mathbf{B}}\}$ is uniquely identifiable and so is $\operatorname{span}\{\widetilde{\mathbf{B}}\}^\perp$ (i.e. $\|\mathbf{P}_{\operatorname{span}\{\widehat{\mathbf{B}}_{k_t}^t\}} - \mathbf{P}_{\operatorname{span}\{\widetilde{\mathbf{B}}\}}\| = \|\widehat{\mathbf{B}}_{k_t}^t(\widehat{\mathbf{B}}_{k_t}^t)^T - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T\| = \|(\mathbf{I}_p - \widetilde{\mathbf{B}}\widetilde{\mathbf{B}}^T) - (\mathbf{I}_p - \widehat{\mathbf{B}}_{k_t}^t(\widehat{\mathbf{B}}_{k_t}^t)^T)\| = \|\mathbf{P}_{\operatorname{span}\{\widetilde{\mathbf{B}}\}^\perp} - \mathbf{P}_{\operatorname{span}\{\widehat{\mathbf{B}}_{k_t}^t\}^\perp}\|$). Thus, all three conditions are met and the result is obtained. □

Proof of Theorem 27. Let $(t_j)_{j=1,\dots,m_n}$ be an i.i.d. sample from F_T and write

$$\begin{aligned} |L_{n,\mathcal{F}}(\mathbf{V}) - L_\mathcal{F}(\mathbf{V})| = &\left| \frac{1}{m_n} \sum_{j=1}^{m_n} \left(L_n^*(\mathbf{V}, f_{t_j}) - L^*(\mathbf{V}, f_{t_j}) \right) \right| \\ &+ \left| \frac{1}{m_n} \sum_{j=1}^{m_n} \left(L^*(\mathbf{V}, f_{t_j}) - \mathbb{E}_{t\sim F_T}(L^*(\mathbf{V}, f_t)) \right) \right| \end{aligned} \tag{4.47}$$

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 (4.47) and let $\delta > 0$. By Markov's and

triangle inequalities and Fatou's Lemma,

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

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

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

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

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

since by Theorem 25 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 (4.47) we apply Theorem 37 with the function $L^*(\mathbf{V},f_{t_j})$. 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 (4.47), $Z_j(\mathbf{V}) = L^*(\mathbf{V},f_{t_j}) < 4M^2$ 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 20 and the parameter space $\mathcal{S}(p,q)$ is compact. Then by Theorem 37,

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

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 $n\to\infty$. □

Proof of Theorem 28. The proof is directly analogous to the proof of Theorem 26. The uniform convergence of the target function $L_{n,\mathcal{F}}(\mathbf{V})$ is obtained by Theorem 27. The minimizer over $Gr(p,q)$ and its uniqueness derive from Theorem 22. □

4.5 Simulations

4.5.1 Simulation Study: Influence of m_n on ecve

In this section we study, via a sumlation study, the influence of m_n, i.e. the number of functions of the ensemble $\mathcal{F}$ used given in (4.21), on the accuracy of ensemble conditional

variance estimation. In Theorem 27 and 28 the rate of $m_n \to \infty$ is unspecified. Therefore we consider the 2-dimensional regression model

$$Y = (\mathbf{b}_2^T \mathbf{X}) + (0.5 + (\mathbf{b}_1^T \mathbf{X})^2)\epsilon, \tag{4.48}$$

where $p = 10$, $k = 2$, $\mathbf{X} \sim N(0, I_{10})$, $\epsilon \sim N(0,1)$ 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 $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2)$.

The sample size is $n = 300$ and we vary m over a range of values, $m \in \{4, 8, 10, 26, 50, 76, 100\}$ for the indicator, characteristic, monomial, and Box-Cox ensemble. That is

- Indicator ensemble: $\mathcal{F}_{m,\text{Indicator}} = \{1_{\{x \geq q_j\}} : j = 1, \ldots, m\}$, where q_j is the $j/(m+1)$th empirical quantile of $(Y_i)_{i=1,\ldots,n}$
- Fourier ensemble: $\mathcal{F}_{m,\text{Fourier}} = \{\sin(jx) : j = 1, \ldots, m/2\} \cup \{\cos(jx) : j = 1, \ldots, m/2\}$
- Monomial ensemble: $\mathcal{F}_{m,\text{Monom}} = \{x^j : j = 1, \ldots, m\}$
- BoxCox ensemble: $\mathcal{F}_{m,\text{BoxCox}} = \{(x^{t_j} - 1)/t_j : t_j = 0.1 + 2(j-1)/(m-1), j = 1, \ldots, m-1\} \cup \{\log(x)\}$.

For each ensemble we form the ensemble conditional variance estimator and the weighted version (see Section 2.2.2). For further comparison the main competitor *csMAVE* is also included. The results of 100 replication for each method and each m are displayed in Figure 4.1. For the Fourier basis fewer basis function give the best performance, the indicator and BoxCox ensemble is quite robust against varying m, and for the monomial ensemble the results get rapidly worse if m is increased. Further the weighted version improves the accuracy for all ensembles and $\mathcal{F}_{4,\text{Fourier\_weighted}}$, $\mathcal{F}_{8,\text{Indicator\_weighted}}$, $\mathcal{F}_{4,\text{BoxCox\_weighted}}$ are on par or slightly more accurate than *csMAVE*.

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

$$m_n = \begin{cases} \lceil \log(n) \rceil, \text{if} \quad \lceil \log(n) \rceil \quad \text{even} \\ \lceil \log(n) \rceil + 1, \text{if} \quad \lceil \log(n) \rceil \quad \text{odd} \end{cases} \tag{4.49}$$

in the following simulations in Section 4.5.

4.5.2 Simulation Study: Demonstrating consistency

We continue by exploring the consistency of the *conditional variance estimator* (CVE) and *ensemble conditional variance estimator* (ECVE) through a simulation study using the same model (4.48).

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.

- $\mathcal{S}_{Y|\mathbf{X}}$
  - I: **Fourier** is ECVE: $\mathcal{F}_{m_n,\text{Fourier}} = \{\sin(jx) : j = 1, \ldots, m_n/2\} \cup \{\cos(jx) : j = 1, \ldots, m_n/2\}$

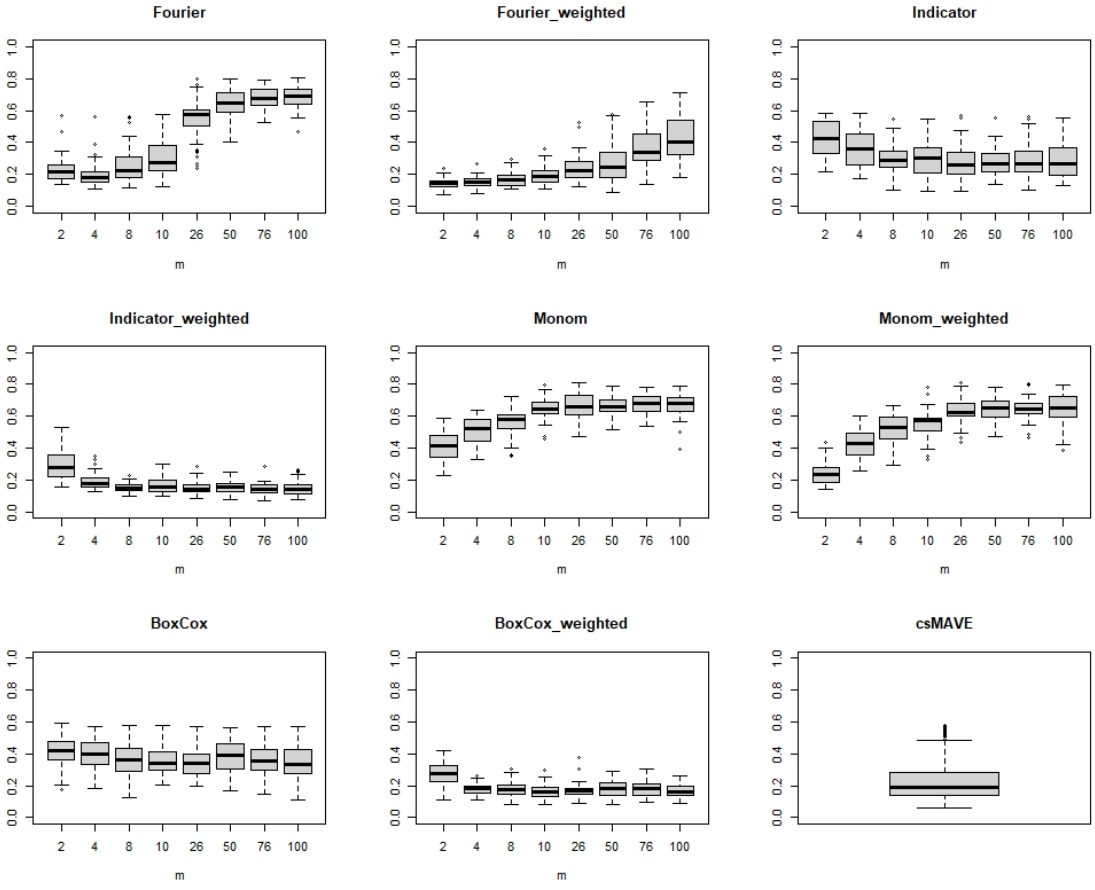


Figure 4.1: Box plots of 100 replications of estimation error from model (4.48) with $n = 300$ over $|\mathcal{F}| = m = (2, 4, 8, 10, 26, 50, 76, 100)$ for the different ensembles

II: **Indicator** is ECVE: $\mathcal{F}_{m_n,\text{Indicator}} = \{1_{\{x \geq q_j\}} : j = 1, \ldots, m_n\}$, where q_j is the $j/(m_n+1)$th empirical quantile of $(Y_i)_{i=1,\ldots,n}$

III: **Monom** is ECVE: $\mathcal{F}_{m_n,\text{Monom}} = \{x^j : j = 1, \ldots, m_n\}$

IV: **BoxCox** is ECVE: $\mathcal{F}_{m_n,\text{BoxCox}} = \{(x^{t_j} - 1)/t_j : t_j = 0.1 + 2(j-1)/(m-1), j = 1, \ldots, m_n - 1\} \cup \{\log(x)\}$.

V: **csMAVE** from [WX08]

- $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$

VI: **cve** of Chapter 2

VII: **meanMAVE** from [XTLZ02]

The simulation is performed as follows. For a given sample size n, 100 i.i.d samples $(Y_i, \mathbf{X}_i^T)_{i=1,\ldots,n}$ are drawn from (4.48). For the first five methods we set $k = 2$ and estimate $\mathbf{B} \in \mathbb{R}^{10\times 2}$, and for the last two, we set $k = 1$ and estimate $\mathbf{b}_2 \in \mathbb{R}^{10\times 1}$ from sample $j = 1, \ldots, 100$ and $\text{err}_{j,n} = \|\hat{\mathbf{B}}\hat{\mathbf{B}}^T - \mathbf{B}\mathbf{B}^T\|/(2k)^{1/2}$ is calculated. This is repeated for sample sizes $n = 100, 200, 400, 600, 800, 1000$. Figure 4.2 displays the distribution of $err_{j,n}$ for increasing n for the seven methods. The plots indicate that as the sample size increases all methods, except the ECVE with Monomial and BoxCox ensemble, yield estimates that are increasingly accurate for their target. In particular, the *central subspace* methods Fourier, Indicator ensemble, and `csMAVE` estimate $\mathcal{S}_{Y|\mathbf{X}} = \text{span}\{\mathbf{B}\}$ consistently and both *mean subspace* methods estimate $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})} = \text{span}\{\mathbf{b}_2\}$ consistently.

4.5.3 Simulations to evaluate estimation accuracy

In this section the performance of the ECVE is studied in simulations. We consider seven models, (M1-M7) defined in Table 4.1, three different sample sizes $\{100, 200, 400\}$, and three different distributions of the predictor vector $\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Z} \in \mathbb{R}^p$, where $\mathbf{\Sigma} = (\Sigma_{ij})_{i,j=1,\ldots,p}$, $\Sigma_{i,j} = 0.5^{|i-j|}$. Throughout, $p = 10$, $\mathbf{B}$ are the first k columns of $\mathbf{I}_p$, and $\epsilon \sim N(0,1)$ independent of $\mathbf{X}$. As in [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 2.2.2] `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 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` or in the `R`-package `CVarE`.

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 $err = \|\mathbf{P}_{\mathbf{B}} - \mathbf{P}_{\hat{\mathbf{B}}}\|_2/\sqrt{2k} \in [0,1]$, 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

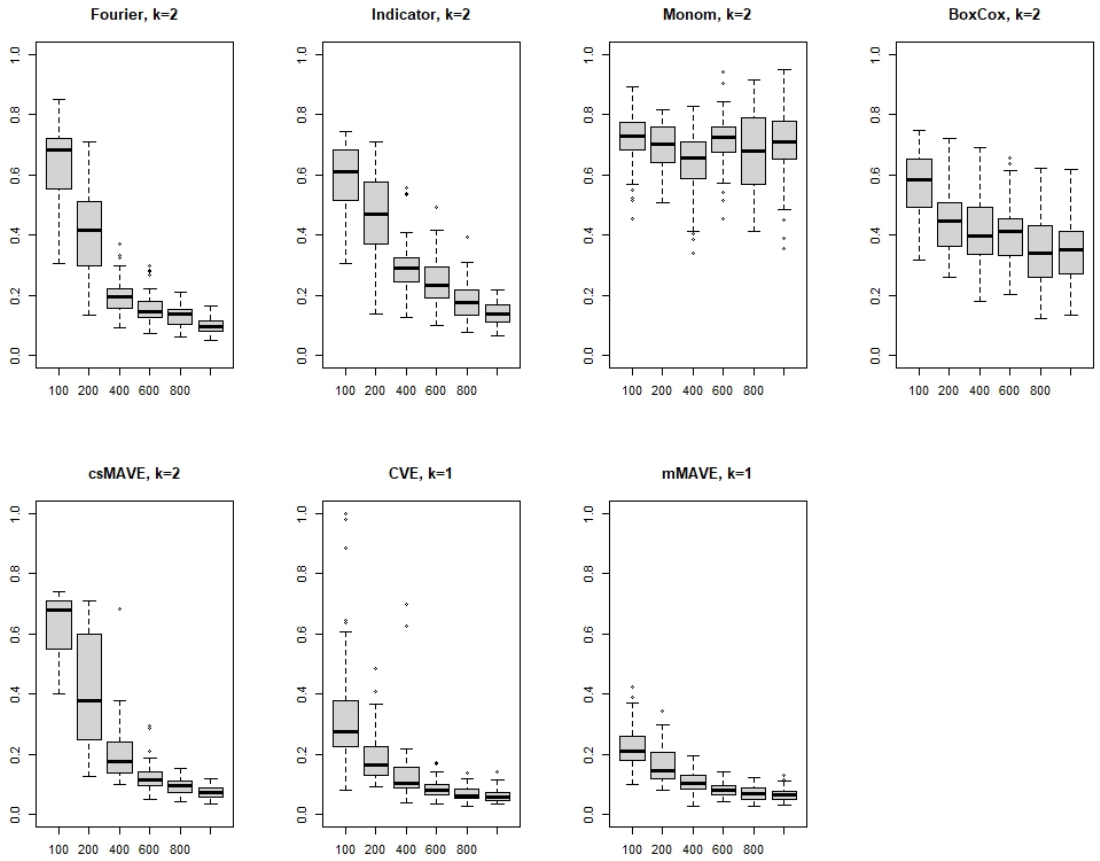


Figure 4.2: Estimation error distribution of model (4.48) plotted over $n = (100, 200, 400, 600, 800, 1000)$ for the seven (I-VII) methods

Table 4.1: 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	$Y = \frac{\mathbf{b}_1^T\mathbf{X}}{0.5+(1.5+\mathbf{b}_2^T\mathbf{X})^2} + (\lvert\mathbf{b}_1^T\mathbf{X}\rvert + (\mathbf{b}_2^T\mathbf{X})^2 + 0.5)\epsilon$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$	2
M5	$Y = \mathbf{b}_3^T\mathbf{X} + \sin(\mathbf{b}_1^T\mathbf{X}(\mathbf{b}_2^T\mathbf{X})^2)\epsilon$	$\text{span}\{\mathbf{b}_3\}$	$\text{span}\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$	3
M6	$Y = 0.5(\mathbf{b}_1^T\mathbf{X})^2\epsilon$	$\text{span}\{\mathbf{0}\}$	$\text{span}\{\mathbf{b}_1\}$	1
M7	$Y = \cos(\mathbf{b}_1^T\mathbf{X} - \pi) + \cos(2\mathbf{b}_1^T\mathbf{X})\epsilon$	$\text{span}\{\mathbf{b}_1\}$	$\text{span}\{\mathbf{b}_1\}$	1

$\sqrt{2k}$ 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 4.2-4.8. 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 4.2. 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 4.3 we see that `csMAVE` is the best performing method. M3 is the same as model (4.48) and here the mean subspace is a proper subset of the central subspace. In Table 4.4 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 4.5 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 4.6 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 4.7, 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 4.8, we see that all ensemble conditional variance estimators clearly outperform `csMAVE`.

4.6 Data Analysis

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 centres), `rad` (index of

Table 4.2: 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 4.3: 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)

Table 4.4: 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 4.5: 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)

Table 4.6: 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 4.7: 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 4.8: 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)

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 4.9 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].

Table 4.9: 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

Moreover, we plot the standardized response `medv` against the reduced `Fourier` and `csMAVE` predictors, $\mathbf{B}^T\mathbf{X}$, in Figure 4.3. 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, $|\widehat{\mathbf{B}}^T_{\texttt{Fourier}}\widehat{\mathbf{B}}_{\texttt{csMAVE}}| = 0.786$. The coefficients of the reductions, $\widehat{\mathbf{B}}_{\texttt{Fourier}}$ and $\widehat{\mathbf{B}}_{\texttt{csMAVE}}$, are reported in Table 4.10. For the `Fourier` ensemble, the variables `rm` and `lstat` have the highest influence on the target variable `medv`. This agrees with the analysis in [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 4.10: Rounded coefficients of the estimated reductions for $\widehat{\mathbf{B}}_{\texttt{Fourier}}$ and $\widehat{\mathbf{B}}_{\texttt{csMAVE}}$ 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

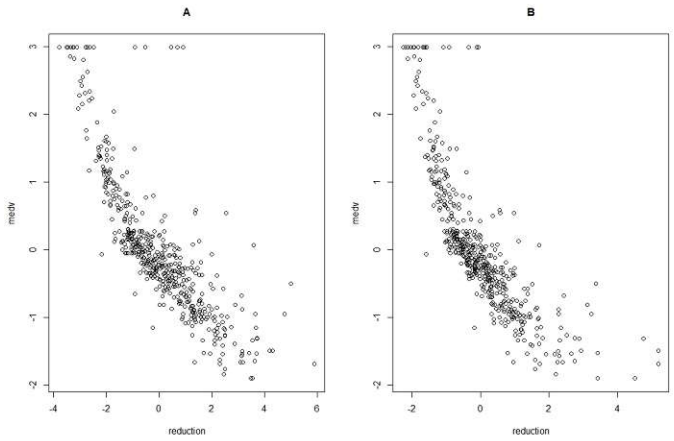


Figure 4.3: Panel A: Y vs. $\widehat{\mathbf{B}}^T_{\texttt{Fourier}}\mathbf{X}$. Panel B: Y vs. $\widehat{\mathbf{B}}^T_{\texttt{csMAVE}}\mathbf{X}$

4.7 Discussion

In this chapter, we extended the *mean subspace* conditional variance estimation (CVE) to the ensemble conditional variance estimation (ECVE), which exhaustively estimates the *central subspace*, by applying the ensemble device introduced by [YL11]. In Section 4.4 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 4.5.3. 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 independence of all quantities in the calculation. The dimension can be estimated via cross-validation, as in [WX08], or information criteria.

5 Conclusion and perspectives for future work

In this thesis, three novel sufficient dimension reduction methods, *conditional variance estimation* (CVE) in Chapter 2, *neural net sufficient dimension reduction* (NN − SDR) in Chapter 3, and *ensemble conditional variance estimation* (ECVE) in Chapter 4 were introduced. The first two are estimators for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$ and the latter is an estimator for the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

The estimators CVE and ECVE are shown to be consistent estimators for their target under relatively mild regularity assumptions in Sections 2.5 and 4.4 with the main theoretical work presented in Section 4.4.1. Moreover, the performance and estimation accuracy of CVE, NN − SDR, and ECVE are explored via simulations and data applications. All three are shown to be competitive against the state-of-the-art sufficient dimension reduction method *minimum average variance estimation* MAVE for the mean subspace and its extension *central subspace minimum average variance estimation* to the central subspace. The simulations also indicate that CVE and ECVE can yield substantial benefits compared to MAVE and central subspace MAVE in terms of estimation accuracy in certain models. In particular, for CVE, we consider model M2 in Section 2.7.2 where the predictors have a bimodal distribution, and, for ECVE, model M7 in Section 4.5.3 with cyclical heteroskedasticity. Both estimators are implemented in the `R` package `CVarE`.

Possible future work includes further theoretical assessments, e.g. computing the rate of convergence, asymptotic distributions of the estimators and a consistent approach to estimate the structural dimension k. The former two are especially interesting to broaden the understanding and comparison of the proposed estimators to established methods like MAVE, for which the rate is known. Moreover, the circumstances where CVE or ECVE enjoy a substantial advantage in terms of accuracy could not be fully characterised yet but would be of practical importance. The proposed methods could probably also be further tuned to increase estimation accuracy by techniques like adaptive bandwidths or data preprocessing via screening methods to increase robustness and performance, as is done for the MAVE estimator in the `R` package `MAVE`. This can be seen in M6 in Section 2.7.2, where the results of MAVE implemented in the highly tuned and consistently updated `MAVE` package differ substantially from the results obtained by the `R` code of MAVE published in [Li18]. Another interesting line of future research is to apply CVE and ECVE to regressions where $n < p$ since formally the algorithm for both does not require the inversion of any matrix and can be applied in such settings. Furthermore, the author hopes that the novel estimation idea underpinning both estimators via identifying the reduction through the orthogonal complement, thus circumventing the inversion of the covariance matrix of the predictors, can be transferred to other estimation techniques.

NN − SDR combines the classic *sufficient dimension reduction* approach with neural nets

to form a novel estimator for the mean subspace. In simulations and data examples it is shown that NN − SDR is competitive in terms of estimation and prediction accuracy with MAVE and CVE for data sizes usually considered in the classic *sufficient dimension reduction* literature. Further, NN − SDR has the substantial advantage that due to its usage of neural nets it can also be efficiently applied to regression problems with big sample sizes where MAVE and CVE are infeasible due to the computational costs. Especially nowadays problems with huge sample size and predictor dimensions are frequently encountered in applications, raising the need for *sufficient dimension reduction methods* that can be used without the computing power of scientific server clusters. For NN − SDR estimation no consistency proof is presented due to the theoretical challenges involving neural nets. Nevertheless, simulations in Sections 3.6 and especially 3.7 indicate that NN − SDR behaves as a consistent estimator for the mean subspace $\mathcal{S}_{\mathbb{E}(Y|\mathbf{X})}$. Future lines of research include the development of statistical theory for the NN − SDR estimator, e.g. showing the consistency and other asymptotic results as mentioned for CVE and ECVE. Moreover, the generalization of NN − SDR for the central subspace would also be of considerable interest. Another interesting line of future work is to analyze CVE and NN − SDR in the presence of collinearity as the simulations in Section 3.8.2.1 indicate that they are quite robust.

All three proposed methods assume the predictors to be continuous. The incorporation of categorical predictors would be of considerable interest for real data applications. Especially for NN − SDR the data applications in Section 3.8.3 point to the direction that these methods can also handle categorical variables.

Bibliography

[AAB+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[AKG19] Mahmood Akhtar, Moritz Kraemer, and Lauren Gardner. A dynamic neural network model for predicting risk of zika in real time. *BMC Medicine*, 17:171, 09 2019.

[AL91] E. Ambikairajah and S. Lennon. Neural networks for speech recognition. In Michael F. McTear and Norman Creaney, editors, *AI and Cognitive Science '90*, pages 163–177, London, 1991. Springer London.

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

[AT20] JJ Allaire and Yuan Tang. *tensorflow: R Interface to 'TensorFlow'*, 2020. R package version 2.2.0.

[BC01] Efstathia Bura and R. Dennis Cook. Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 63(2):393–410, 2001.

[BDF16] Efstathia Bura, Sabrina Duarte, and Liliana Forzani. Sufficient reductions in regressions with exponential family inverse predictors. *Journal of the American Statistical Association*, 111(515):1313–1329, 2016.

[BF15] Efstathia Bura and Liliana Forzani. Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association*, 110(509):420–434, 2015.

[BJM06] Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Comment. *Statist. Sci.*, 21(3):341–346, 08 2006.

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

[Bot98] Léon Bottou. Online learning and stochastic approximations, 1998.

[Bri12] David Brillinger. A generalized linear model with "gaussian" regressor variables. *A Festschrift for Erich L. Lehmann*, 11 2012.

[BY11] E. Bura and J. Yang. Dimension estimation in sufficient dimension reduction: A unifying approach. *Journal of Multivariate Analysis*, 102(1):130–142, 2011.

[CC02] Francesca Chiaromonte and R. Dennis Cook. Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 54:768–795, 01 2002.

[CF08] R. Dennis Cook and Liliana Forzani. Principal fitted components for dimension reduction in regression. *Statistical Science*, 23(4):485–501, 2008.

[CF09] R. Dennis Cook and Liliana Forzani. Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, 104(485):197–208, 3 2009.

[Chi94] Yasuko Chikuse. *Invariant measures on Stiefel manifolds with applications to multivariate analysis*, volume Volume 24 of *Lecture Notes–Monograph Series*, pages 177–193. Institute of Mathematical Statistics, Hayward, CA, 1994.

[Chi03] Yasuko Chikuse. *Statistics on Special Manifolds*. Springer-Verlag New York, New York, 2003.

[CHS81] Stamatis Cambanis, Steel Huang, and Gordon Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368 – 385, 1981.

[CL95] R. J. Carroll and Ker-Chau Li. Binary regressors in dimneiosn reduction models: A new look at treatment comparison. *Statistica Sinica*, 5(2):667–688, 1995.

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

[CL04] R. Dennis Cook and Bing Li. Determining the dimension of iterative hessian transformation. *Ann. Statist.*, 32(6):2501–2531, 12 2004.

[Coo98a] R. Dennis Cook. Principal hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94, 1998.

[Coo98b] R. Dennis Cook. Principal hessian directions revisited. *Journal of the American Statistical Association*, 93(441):84–94, 1998.

[Coo98c] R. Dennis Cook. *Regression Graphics: Ideas for studying regressions through graphics*. Wiley, New York, 1998.

[Coo00] R. Dennis Cook. Save: A method for dimension reduction and graphics in regression. *Communications in Statistics - Theory and Methods*, 29:2109–2121, 09 2000.

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

[Coo18] R. Dennis Cook. Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application*, 5(1):533–559, 2018.

[CW91] R. Dennis Cook and Sanford Weisberg. Sliced inverse regression for dimension reduction: Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991.

[CY01] R. Dennis Cook and X. Yin. Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics*, 43:147–199, 06 2001.

[DF84] Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 09 1984.

[Eat86] Morris L. Eaton. A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272 – 276, 1986.

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

[Fas02] Beat Fasel. Robust face analysis using convolutional neural networks. volume 2, pages 40 – 43 vol.2, 02 2002.

[FB21a] Lukas Fertl and Efstathia Bura. Conditional Variance Estimator for Sufficient Dimension Reduction. *arXiv:2102.08782 [math, stat]*, February 2021. arXiv: 2102.08782.

[FB21b] Lukas Fertl and Efstathia Bura. Ensemble Conditional Variance Estimator for Sufficient Dimension Reduction. *arXiv:2102.13435 [stat]*, February 2021. arXiv: 2102.13435.

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

[FS81] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823, 1981.

[FTAW20] Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local Linear Forests. *arXiv:1807.11408 [cs, econ, math, stat]*, September 2020. arXiv: 1807.11408.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[Geo12] Geoffrey Hinton with Nitish Srivastava and Kevin Swersky. Neural Networks for Machine Learning Lecture 6a - Overview of mini-batch gradient descent, 2012.

[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.

[GKC19] Benyamin Ghojogh, F. Karray, and Mark Crowley. Eigenvalue and generalized eigenvalue problems: Tutorial. *ArXiv*, abs/1903.11240, 2019.

[Gur97] Kevin Gurney. *An Introduction to Neural Networks*. Taylor & Francis, Inc., USA, 1997.

[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.

[HHI93] Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993.

[HJ17] J. Hoffmann-Jørgensen. *Probability with a view towards statistics*. Chapman and Hall, 01 2017.

[HL93] Peter Hall and Ker-Chau Li. On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, 21(2):867–889, 1993.

[Hor91] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[HR17] Trevor Hastie and Tibshirani Robert. *mda: Mixture and Flexible Discriminant Analysis*, 2017. S original by Trevor Hastie & Robert Tibshirani. Original R port by Friedrich Leisch and Kurt Hornik and Brian D. Ripley. R package version 0.4-10.

[HS89] Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989.

[HSS08] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220, 06 2008.

[HT90] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Monographs on Statistics and Applied Probability, Volume 43. Chapman and Hall/CRC, 1990.

[HTFF04] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell.*, 27:83–85, 11 2004.

[Ich93] Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71 – 120, 1993.

[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.

[JWHT14] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014.

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

[KF21] Daniel Kapla and Lukas Fertl. *CVarE: Conditional Variance Estimator for Sufficient Dimension Reduction*, 2021. R package version 1.1.

[KFB21] Daniel Kapla, Lukas Fertl, and Efstathia Bura. Fusing sufficient dimension reduction with neural networks, 2021. arXiv: 2104.10009.

[KG12] Sigbert Klinke and Janet Grassmann. *Projection Pursuit Regression*, chapter 16, pages 471–496. John Wiley and Sons, Ltd, 2012.

[KJB08] V. R. V. Krishnan, A. Jayakumar, and A. P. Babu. Speech recognition of isolated malayalam words using wavelet features and artificial neural network. In *4th IEEE International Symposium on Electronic Design, Test and Applications (delta 2008)*, pages 240–243, 2008.

[Kol08] Michael Kolonk. *Stochastische Simulation*. Vieweg+Teubner Verlag, 2008.

[KR15] Thomas Kämpke and Franz Josef Radermacher. *The Generalized Inverse of Distribution Functions*, pages 9–28. Springer International Publishing, Cham, 2015.

[Kra91] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991.

[Kus14] Norbert Kusolitsch. *Maß- und Wahrscheinlichkeitstheorie*. Springer Lehrbuch. Springer Spektrum, 2014.

[KW19] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *arXiv e-prints*, page arXiv:1906.02691, June 2019.

[LB98] Y. LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. 1998.

[LD89] Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):1009–1052, 1989.

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

[Li92] Ker-Chau Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039, 1992.

[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.

[LLC13] Kuang-Yao Lee, Bing Li, and Francesca Chiaromonte. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Statist.*, 41(1):221–249, 02 2013.

[LW07a] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[LW07b] Bing Li and Shaoli Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008, 2007.

[LZC05a] Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, 33(4):1580 – 1616, 2005.

[LZC05b] Bing Li, Hongyuan Zha, and Francesca Chiaromonte. Contour regression: A general approach to dimension reduction. *Ann. Statist.*, 33(4):1580–1616, 2005.

[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.

[MP43] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[MRT12] M. Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. In *Adaptive computation and machine learning*, 2012.

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

[Nad05] Saralees Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005.

[NW06] J. Nocedal and S. Wright. *Line Search Methods*, pages 30–65. Springer New York, New York, NY, 2006.

[Par61] E Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1961.

[PHPP18] M.E. Paoletti, J.M. Haut, J. Plaza, and A. Plaza. A new deep convolutional neural network for fast hyperspectral image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:120–147, 2018. Deep Learning RS Data.

[Rad15] Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266 – 282, 2015.

[SCW07] Yongwu Shao, R. Dennis Cook, and Sanford Weisberg. Marginal tests with sliced average variance estimation. *Biometrika*, 94(2):285–296, 2007.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[Sil86] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[SL18] Lukas Steinberger and Hannes Leeb. On conditional moments of high-dimensional random vectors given lower-dimensional projections. *Bernoulli*, 24(1):565–591, 2018.

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

[Spe91] Donald Specht. A general regression neural network. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 2:568–76, 02 1991.

[SS18] B. Scholkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive Computation and Machine Learning. MIT Press, 2018.

[Sto74] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

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

[Tuz20] Alexey A. Tuzhilin. Lectures on hausdorff and gromov-hausdorff distance geometry, 2020.

[Woo08] Simon Wood. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B*, 70:495–518, 07 2008.

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

[WY10] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142, 12 2010.

[WY13] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142:397–434, 2013.

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

[Xia07] Yingcun Xia. A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.*, 35(6):2654–2690, 12 2007.

[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.

[Yin] Xiangrong Yin. *Sufficient Dimension Reduction in Regression*, pages 257–273.

[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. Dennis Cook. Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99:1733–1757, 09 2008.

[ZGD+17] Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, September 2017.

[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.

Curriculum Vitae

Personal data

Name	**Lukas Fertl**
Date of birth	30.08.1992
Birth place	Vienna
Nationality	Austria
Email	lukas.fertl@tuwien.ac.at, lukas.fertl@chello.at
Homepage	http://astat.tuwien.ac.at/fertl/

Education

10/2017 – 03/2021	PhD position in statistics at TU Vienna at department of applied statistics (ASTAT)
10/2015–10/2019	Msc in financial and actuarial mathematics at TU Vienna
10/2015–06/2017	Msc in statistics and mathematics in economics at TU Vienna
10/2012–10/2015	Bsc statistics and mathematics in economics at TU Vienna
10/10–06/11	civil service at Vienna hospital Semmelweis
2002–2010	diploma with honors of Austrian type of secondary school (high school) with focus on natural sciences and mathematics, Real-gymnasium Albertus Magnus school, 1180 Vienna
1998–2002	primary school: Albertus Magnus school, 1180 Vienna

---

Publications

L. Fertl and E. Bura: *Conditional Variance Estimator for Sufficient Dimension Reduction, arXiv:2102.08782 [math, stat]*
L. Fertl and E. Bura: *Ensemble Conditional Variance Estimator for Sufficient Dimension Reduction, arXiv:2102.13435 [stat]*

D. Kapla, L. Fertl and E. Bura: *Fusing Sufficient Dimension Reduction with Neural Networks*

Wien, am June 15, 2021

Lukas Fertl