On dimension reduction in conditional dependence models

Thomas Nagler[1], Gerda Claeskens[2], Irène Gijbels[3]

[1]Department of Statistics, LMU Munich, Ludwigstraße 33, Munich, 80539, Germany.
[2]ORStat and Leuven Statistics Research Center, KU Leuven, Naamsestraat 69, box 3555, Leuven, 3000, Belgium.
[3]Department of Mathematics and Leuven Statistics Research Center, KU Leuven, Celestijnenlaan 200B, box 2400, Leuven, 3001, Belgium.

Contributing authors: t.nagler@lmu.de; gerda.claeskens@kuleuven.be; irene.gijbels@kuleuven.be;

Abstract

Inference of the conditional dependence structure is challenging when many covariates are present. In numerous applications, only a low-dimensional projection of the covariates influences the conditional distribution. The smallest subspace that captures this effect is called the central subspace in the literature. We show that inference of the central subspace of a vector random variable $\boldsymbol{Y}$ conditioned on a vector of covariates $\boldsymbol{X}$ can be separated into inference of the marginal central subspaces of the components of $\boldsymbol{Y}$ conditioned on $\boldsymbol{X}$ and on the copula central subspace, that we define in this paper. Further discussion addresses sufficient dimension reduction subspaces for conditional association measures. An adaptive nonparametric method is introduced for estimating the central dependence subspaces, achieving parametric convergence rates under mild conditions. Simulation studies illustrate the practical performance of the proposed approach.

Keywords: dimension, reduction, conditional, copula, dependence, nonparametric

1 Introduction

Dependence among the components of a random vector $\boldsymbol{Y}$ of length q is an important concept in statistics and many applied disciplines. Medical scientists analyze the dependence in chemical profiles of tumor cells (Helmlinger et al. 1997), economists study the dependence between economic indicators (Rodriguez 2007), hydrologists model the dependence between flood characteristics (Zhang and Singh 2006), etc.

Often one also wants to investigate or control for the effect of other variables $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ on the dependence in $\boldsymbol{Y}$. Continuing the examples from the fist paragraph, the medical scientist may be interested how the dependence is influenced by the patient's age or the time passed since outbreak of cancer, the economists may want to control for a country's size and demography when analyzing the dependence between economic indicators, and the hydrologists could ask for the influence of geographical properties on the dependence between flood characteristics.

Such questions can be answered by conditional dependence models. We approach this via modeling the conditional copula of $\boldsymbol{Y}$ given $\boldsymbol{X}$. This is an active field of research. Several inference methods from conditional copula models have been investigated. Patton (2001) considered fully parametric models where the copula's parameter can be expressed as as a known, parametric function of the covariates. More flexible semiparametric models were discussed by Acar et al. (2011), Abegaz et al. (2012), Vatter and Chavez-Demoulin (2015), and Fermanian and Lopez (2018). Here, the copula parameter is allowed to exhibit an unknown, non- or semiparametric relationship with the covariates. Fully nonparametric estimators of the conditional copula and association measures were discussed by Veraverbeke et al. (2011) and Gijbels et al. (2011). In many of these contributions the dimension p of the covariate vector $\boldsymbol{X}$ is assumed to be small (often $p = 1$). The reasons are diverse but evident: In parametric and semiparametric models, structural assumptions such as linearity and additivity become more restrictive when p is large. Additionally, issues with identifiability and numerical instability arise. For nonparametric methods, the *curse of dimensionality* deteriorates the accuracy of the estimators; a dimension $p \geq 5$ is often considered too large already.

A popular solution to issues with high dimensional covariates are techniques for sufficient dimension reduction. The goal of such methods is to find a projection of the covariate vector $\boldsymbol{X}$ onto a space of dimension $d \ll p$ such that all of the relevant information in $\boldsymbol{X}$ is preserved. In regression problems, the theory of *central subspaces* (Cook 1994, 1998; Cook and Li 2002) serves as the foundation of a broad variety of methods. Li et al. (2011) developed a principal support vector machine approach for use in linear and nonlinear sufficient dimension reduction. Haffenden and Artemiou (2024) used the sliced inverse mean difference for dimension reduction and Zhang et al. (2024) studied dimension reduction for Fréchet regression, to give just a few examples.

In this paper we study dimension reduction methods in conditional dependence models and find that this is closely related to dimension reduction in regression problems. The main definitions and properties of these are briefly revised in Section 3, after introducing the copula notation in Section 2. Novel concepts, such as a central copula subspace, we define in Section 4 and we study some of its theoretical properties. Further, we introduce and study sufficient dimension reduction in the context of conditional association measures and concordance measures in Section 5 and Section 6. Estimation of the central subspaces is explained in Section 7. Section 8 contains simulation results. Longer proofs are collected in the Appendix.

2 Notation, definitions and background on copulas

The most general tool to statistically model the dependence in multivariate distributions is the *copula*. According to Sklar's theorem (Sklar 1959), the joint distribution F of the random variables $Y_1 \sim F_1$, ..., $Y_q \sim F_q$ can be expressed as

$$F(\boldsymbol{y}) = C\big\{F_1(y_1), \dots, F_q(y_q)\big\}, \qquad \text{for all } \boldsymbol{y} \in \mathbb{R}^q, \tag{2.1}$$

where the function $C\colon [0,1]^q \to [0,1]$ is called the copula of the random vector $\boldsymbol{Y} = (Y_1, \dots, Y_q)^\top$. By its definition, a copula is a joint distribution function of uniformly distributed random variables. If $Y_1, \dots, Y_q$ are continuous random variables, what we shall assume from here on, the copula C is unique. More specifically, C is the multivariate distribution function of the random vector $\boldsymbol{U} = (U_1, \dots, U_q)^\top = (F_1(Y_1), \dots, F_q(Y_q))^\top$. Hence, $C(\boldsymbol{u}) = P\big(U_1 \le u_1, \dots, U_q \le u_q\big)$, for all $\boldsymbol{u} \in [0,1]^q$. The reverse is also true, for arbitrary marginals $F_1, \dots, F_q$ and copula C, (2.1) is a valid joint distribution. Sklar's theorem shows that any joint distribution can be decomposed into the marginal distributions and the copula. Hence, the copula captures all of the non-marginal behavior of the random vector, i.e., the dependence between the random variables.

Patton (2001) stated a version of Sklar's theorem that includes a conditioning on $\boldsymbol{X}$. Let $F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}) = P(\boldsymbol{Y} \le \boldsymbol{y}|\boldsymbol{X})$ be the distribution of a random vector $\boldsymbol{Y}$ conditional on $\boldsymbol{X}$. Similarly, denote $F_{Y_j|\boldsymbol{X}}(y_j) = P(Y_j \le y_j|\boldsymbol{X})$, $j = 1, \dots, q$ as the conditional marginal distributions. Note that, due to their dependence on $\boldsymbol{X}$, the conditional joint and marginal distributions are random functions. Then for all $\boldsymbol{y} \in \mathbb{R}^q$ and almost surely (*a.s.* from here on),

$$F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}) = C_{\boldsymbol{X}}\big\{F_{Y_1|\boldsymbol{X}}(y_1), \dots, F_{Y_q|\boldsymbol{X}}(y_q)\big\},$$

for some (random) copula function $C_{\boldsymbol{X}}\colon [0,1]^q \to [0,1]$ which is called the conditional copula associated with the random vector $\boldsymbol{Y}|\boldsymbol{X}$. It is the conditional multivariate distribution function of the random vector

$$\boldsymbol{U}_{\boldsymbol{X}} = (U_{1,\boldsymbol{X}}, \dots, U_{q,\boldsymbol{X}})^\top = \big(F_{Y_1|\boldsymbol{X}}(Y_1), \dots, F_{Y_q|\boldsymbol{X}}(Y_q)\big)^\top,$$

conditional on $\boldsymbol{X}$, i.e., for all $\boldsymbol{u} \in [0,1]^q$,

$$C_{\boldsymbol{X}}(\boldsymbol{u}) = P\big(U_{1,\boldsymbol{X}} \le u_1, \dots, U_{q,\boldsymbol{X}} \le u_q|\boldsymbol{X}\big) \quad a.s.$$

3 Background on sufficient dimension reduction in regression problems

First, we revise the needed definitions and concepts before introducing the novel copula dimension reduction methods in Section 4.

Consider a regression problem with (possibly multivariate) response $\boldsymbol{Y} \in \mathbb{R}^q$, $q \in \mathbb{N}$, and covariate vector $\boldsymbol{X} \in \mathbb{R}^p$. We are interested in the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Suppose there exists a projection matrix $\boldsymbol{B} \in \mathbb{R}^{p \times d}$, $d \leq p$, such that for all $\boldsymbol{y} \in \mathbb{R}^q$, $P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{X}) = P(\boldsymbol{Y} \leq \boldsymbol{y} | \boldsymbol{B}^\top \boldsymbol{X}) \quad a.s.$, or more concisely:

$$F_{\boldsymbol{Y}|\boldsymbol{X}} = F_{\boldsymbol{B}^\top \boldsymbol{X}} \quad a.s. \tag{3.1}$$

Then $\boldsymbol{B}^\top \boldsymbol{X}$ contains all the information of $\boldsymbol{X}$ that is relevant for predicting $\boldsymbol{Y}$. The $p \times p$ identity matrix trivially satisfies (3.1), so a matrix $\boldsymbol{B}$ satisfying (3.1) always exists. When such a matrix exists for $d < p$, we can reduce the dimension of the covariate vector without losing any information about the conditional expectation. In other words, knowledge of the d-dimensional vector $\boldsymbol{B}^\top \boldsymbol{X}$ is *sufficient* for predicting Y and we speak of *sufficient dimension reduction*.

If we can find such a matrix for $d \ll p$, we benefit in two ways. Inference of the reduced regression problem becomes much easier than the original one. In particular, the curse of dimensionality inherent in nonparametric regression problems can be mitigated substantially. Further, if $q = 1$, plotting the response variable Y against the reduced predictor $\boldsymbol{B}^\top \boldsymbol{X}$ provides a low-dimensional visualization of their relationship. The initial dimension does not have to be large to make the reduction useful. Even when $p = 5$, the curse of dimensionality drastically deteriorates the quality of nonparametric estimates and the data is difficult to visualize. These problems disappear when the dimension can be reduced to $d \leq 2$ which is typically the goal.

3.1 The central subspace in regression models

It is easy to see that the matrix $\boldsymbol{B}$ is not identifiable. In fact, $\boldsymbol{B}$ could be replaced by any matrix whose columns span the same space as the columns of $\boldsymbol{B}$. Hence, the goal is to identify $\mathcal{S}(\boldsymbol{B})$, the column space of $\boldsymbol{B}$. Whenever $\boldsymbol{B}$ satisfies (3.1), the space $\mathcal{S}(\boldsymbol{B})$ is called a *dimension reduction subspace*. The ultimate goal is to find the smallest dimension reduction subspace, called *central subspace*.

Definition 3.1 (Central subspace) Let $\mathcal{B}_{Y|\boldsymbol{X}} = \{\boldsymbol{B} \in \mathbb{R}^{p \times p} \colon F_{\boldsymbol{Y}|\boldsymbol{X}} = F_{\boldsymbol{B}^\top \boldsymbol{X}} \;\; a.s.\}$. If $\mathcal{S}_{Y|\boldsymbol{X}} = \bigcap_{\boldsymbol{B} \in \mathcal{B}_{Y|\boldsymbol{X}}} \mathcal{S}(\boldsymbol{B})$ is a dimension reduction subspace, it is called the **central subspace**.

The set $\bigcap_{\boldsymbol{B} \in \mathcal{B}_{Y|\boldsymbol{X}}} \mathcal{S}(\boldsymbol{B})$ always contains at least the zero vector. But the central subspace may not exist, namely when $\bigcap_{\boldsymbol{B} \in \mathcal{B}_{Y|\boldsymbol{X}}} \mathcal{S}(\boldsymbol{B})$ is not a dimension reduction subspace. But existence can be guaranteed under rather mild conditions; and whenever the central subspace exists, it is also unique (see, Chiaromonte and Cook 1997; Cook 1998). One simple condition is that the support of $\boldsymbol{X}$ is convex. More complex conditions also applicable to discrete $\boldsymbol{X}$ can be found in Cook (1998).

Even if the central subspace does not exist, a dimension reduction subspace always does. Hence, we can always find a matrix $\boldsymbol{B}$ that satisfies (3.1), which is what is most important in practice. But it is possible that there are multiple dimension reduction subspaces whose intersection is not a dimension reduction subspace. This phenomenon poses a conceptual difficulty, but has only minor relevance in practice. We shall assume throughout that central subspaces exist.

The following examples illustrate the new concepts for readers unfamiliar with them.

Example 3.1 (Single index model) Consider the single index model

$$Y = g(\boldsymbol{\beta}^\top \boldsymbol{X}) + \sigma\epsilon, \quad \epsilon \sim \mathcal{N}(0,1),$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma > 0$ are model parameters and $g\colon \mathbb{R} \to \mathbb{R}$ a possibly unknown function. The classical linear model is recovered for $g(x) = x$. Equivalently, we can write $F_{Y|\boldsymbol{X}} = \Phi_{g(\boldsymbol{\beta}^\top \boldsymbol{X}),\sigma^2}$, where Φ_{μ,σ^2} denotes the normal distribution function with mean μ and variance σ^2. For any matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d'}$, $1 \le d' \le p$, that has $\boldsymbol{\beta}$ in one of its columns, it holds $F_{\boldsymbol{Y}|\boldsymbol{X}} = F_{\boldsymbol{B}^\top \boldsymbol{X}}$ a.s. Thus, $\mathcal{S}(\boldsymbol{B})$ is a mean dimension reduction subspace. The intersection of all these spaces is $\mathcal{S}(\boldsymbol{\beta})$, which is itself a dimension reduction subspace. Hence, $\mathcal{S}_{Y|\boldsymbol{X}} = \mathcal{S}(\boldsymbol{\beta})$ is the central subspace. If $\boldsymbol{\beta} = \mathbf{0}$, i.e., $Y = g(0) + \sigma\epsilon$, it holds $\mathcal{S}_{Y|\boldsymbol{X}} = \{\mathbf{0}\}$ and the dimension of the central subspace is $d = 0$. □

Example 3.2 (Heteroscedastic multivariate model) Now consider

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{\beta}^\top \boldsymbol{X}) + \boldsymbol{A}(\boldsymbol{\gamma}^\top \boldsymbol{X})\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{q\times q})$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$, $\boldsymbol{g}\colon \mathbb{R} \to \mathbb{R}^q$, $\boldsymbol{A}\colon \mathbb{R} \to \mathbb{R}^{q\times q}$, and $\boldsymbol{I}_{q\times q}$ denotes the identity matrix of dimension $q \times q$. The conditional distribution is $F_{\boldsymbol{Y}|\boldsymbol{X}} = \Phi_{\boldsymbol{g}(\boldsymbol{\beta}^\top \boldsymbol{X}),\boldsymbol{A}(\boldsymbol{\gamma}^\top \boldsymbol{X})\boldsymbol{A}(\boldsymbol{\gamma}^\top \boldsymbol{X})^\top}$, where $\Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$ is the multivariate Gaussian cumulative distribution function. The corresponding central subspace is spanned by the columns of $\boldsymbol{B} = (\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^{p\times 2}$. □

3.2 The central mean subspace in regression models

In some cases we are only interested in the conditional mean, not the entire distribution. Suppose there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d}$, $d \le p$, such that

$$\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{B}^\top \boldsymbol{X}) \quad a.s. \tag{3.2}$$

For any matrix $\boldsymbol{B}$ that satisfies (3.2), $\mathcal{S}(\boldsymbol{B})$ is called *mean dimension reduction subspace* (Cook 1994). Again, the ultimate goal is to find the smallest mean dimension reduction subspace.

Definition 3.2 (Central mean subspace) Let $\mathcal{B}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \{\boldsymbol{B} \in \mathbb{R}^{p\times p}\colon \mathbf{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathrm{E}(\boldsymbol{Y}|\boldsymbol{B}^\top \boldsymbol{X})\ a.s\}$. If $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \bigcap_{\boldsymbol{B}\in\mathcal{B}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})}} \mathcal{S}(\boldsymbol{B})$ is a mean dimension reduction subspace, it is called the **central mean subspace**.

Also the central mean subspace may not exist. But its existence can be guaranteed under conditions similar to the ones for the central subspace. A sufficient condition is that the support of $\boldsymbol{X}$ is convex (cf., Cook 1998; Cook and Li 2002; Cook and Setodji 2003).

The central mean subspaces in Example 3.1 is easily derived as $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathcal{S}(\boldsymbol{\beta})$, which is equal to the central subspace $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$. Equality holds because the influence of $\boldsymbol{X}$ on $\boldsymbol{Y}$ is completely captured by the conditional mean. In Example 3.2, the central mean subspace is also $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} = \mathcal{S}(\boldsymbol{\beta})$. But now, $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} \neq \mathcal{S}_{Y|\boldsymbol{X}}$, because $\boldsymbol{X}$ also affects another other aspect of the conditional distribution (the covariance).

3.3 Connection between central and central mean subspaces

Since the conditional expectation is a functional of the conditional distribution, it is clear that $\mathcal{S}_{\mathrm{E}(\boldsymbol{Y}|\boldsymbol{X})} \subseteq \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$. The intuitive consequence is that the dimension can possibly be reduced further when only the conditional mean is of concern. Under a location model (the covariate vector $\boldsymbol{X}$ affects the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ only through the mean) the two spaces coincide.

More generally, it is possible to characterize the central subspace as the union of central mean subspaces indexed by a family of functions $\mathcal{G}$. This fact was noticed in a series of papers (Yin and Cook 2002; Zhang and Singh 2006; Zeng and Zhu 2010) and summarized in a general setting by Yin and Li (2011, Theorem 2.1) that for completeness we repeat here:

Proposition 3.1 *Consider a random vector $\boldsymbol{Y} \in \mathbb{R}^q$ and a family of functions $\mathcal{G}$. If*

(i) all $g \in \mathcal{G}$ are measurable maps from $\mathbb{R}^q$ to $\mathbb{R}$ and satisfy $\mathrm{Var}\{g(\boldsymbol{Y})\} < \infty$,

(ii) $\mathcal{G}$ is dense in the set $\{\mathbb{1}(\boldsymbol{Y} \in A)\colon A$ is a Borel set in $\mathbb{R}^q\}$,

it holds that $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} = \mathrm{span}\{S_{\mathrm{E}\{g(\boldsymbol{Y})|\boldsymbol{X}\}}, g \in \mathcal{G}\}$.

Examples of families $\mathcal{G}$ complying with Proposition 3.1 are the collection of indicator functions $\mathbb{1}(\,\cdot \leq \boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{R}^q$, and the set of polynomials (provided all moments exist). Yin and Li (2011, Theorem 2.2) further show that, almost surely, a finite number of functions $g \in \mathcal{G}$ is sufficient to cover $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$. An important implication is: inference for the central subspace can be based on inference of a finite number of central mean subspaces.

4 Copula decompositions in sufficient dimension reduction

Recall from Section 2 that the conditional joint distribution $F_{\boldsymbol{Y}|\boldsymbol{X}}$ of a random vector $\boldsymbol{Y}$ given $\boldsymbol{X}$ can be decomposed into the conditional marginals $F_{1,\boldsymbol{X}}, \dots, F_{q,\boldsymbol{X}}$ and the conditional copula $C_{\boldsymbol{X}}$. This suggests a similar decomposition for the central subspace.

4.1 Copula decomposition of the dimension reduction subspaces

Lemma 4.1 *Suppose a matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d}$, $d \leq p$, satisfies*

$$F_{\boldsymbol{Y}|\boldsymbol{X}} = F_{\boldsymbol{B}^\top \boldsymbol{X}} \quad a.s. \tag{4.1}$$

Then there are matrices $\boldsymbol{B}_j \in \mathbb{R}^{p\times d_j}$, $j = 1, \dots, q$, and $\boldsymbol{B}_C \in \mathbb{R}^{p\times d_C}$ such that $\max\{d_1, \dots, d_q, d_C\} \leq d$ that are sub-matrices of $\boldsymbol{B}$ satisfying

$$\mathrm{span}\{\mathcal{S}(\boldsymbol{B}_1), \dots, \mathcal{S}(\boldsymbol{B}_q), \mathcal{S}(\boldsymbol{B}_C)\} = \mathcal{S}(\boldsymbol{B}), \tag{4.2}$$

$$F_{Y_j|\boldsymbol{X}} = F_{Y_j|\boldsymbol{B}_j^\top \boldsymbol{X}}, \quad \text{for } j = 1, \dots, q, \quad \text{and} \quad C_{\boldsymbol{X}} = C_{\boldsymbol{B}_C^\top \boldsymbol{X}} \quad a.s. \tag{4.3}$$

Proof Set $\boldsymbol{B}_1 = \cdots = \boldsymbol{B}_j = \boldsymbol{B}_C = \boldsymbol{B}$, which trivially satisfies (4.2). Further note that

$$F_{Y_j|\boldsymbol{X}}(\cdot) = F_{\boldsymbol{Y}|\boldsymbol{X}}(\infty, \ldots, \infty, \cdot, \infty, \ldots \infty),$$

where ∞ is in all but the jth position, and $C_{\boldsymbol{X}} = F_{\boldsymbol{Y}|\boldsymbol{X}}\{F^{-1}_{Y_1|\boldsymbol{X}}, \ldots, F^{-1}_{Y_q|\boldsymbol{X}}\}$. Thus, (4.1) implies (4.3). □

Lemma 4.1 implies that any dimension reduction subspace for the joint distribution can be decomposed into dimension reduction subspaces for the marginals and for the copula. The decomposition is generally not disjoint; dimension reduction subspaces for marginals and the copula can overlap. The next result reverses Lemma 4.1: the span of dimension reduction subspaces for marginals and the copula is a dimension reduction subspace for the joint distribution.

Lemma 4.2 *Suppose there are matrices $\boldsymbol{B}_j \in \mathbb{R}^{p \times d_j}$, $j = 1, \ldots, q$, and $\boldsymbol{B}_C \in \mathbb{R}^{p \times d_C}$ with $\max\{d_1, \ldots, d_q, d_C\} \leq p$ satisfying (4.3). Then there is a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times d}$, $d \leq \max\{\sum_{j=1}^q d_j + d_C, p\}$, that satisfies (4.2) and (4.1).*

Proof Set $\bar{\boldsymbol{B}} = (\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q, \boldsymbol{B}_C) \in \mathbb{R}^{p \times (\sum_j d_j + d_C)}$. If $\sum_j d_j + d_C \leq p$, set $\boldsymbol{B} = \bar{\boldsymbol{B}}$, otherwise set $\boldsymbol{B}$ to a matrix derived from $\bar{\boldsymbol{B}}$ by eliminating linearly dependent columns until $\boldsymbol{B} \in \mathbb{R}^{p \times p}$. For this construction, there are full rank matrices $\boldsymbol{A}_j$, $j = 1, \ldots, q$, $\boldsymbol{A}_C \in \mathbb{R}^{p \times p}$ such that $\boldsymbol{A}_j \boldsymbol{B} = (\boldsymbol{B}_j, \ldots)$, and $\boldsymbol{A}_C \boldsymbol{B} = (\boldsymbol{B}_C, \ldots)$. Then the claim follows from $F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}) = C_{\boldsymbol{X}}\left\{F_{Y_1|\boldsymbol{X}}(y_1), \ldots, F_{Y_q|\boldsymbol{X}}(y_q)\right\}$ *a.s.*, and the fact that conditional probabilities are invariant to full rank transformations of the conditioning vector. □

4.2 The central copula subspace

Lemmas 4.1 and 4.2 can be seen as the dimension reduction version of Sklar's theorem. A natural question is whether similar results can be derived for central subspaces. Paralleling the definitions in Section 3.1, we define a *copula dimension reduction subspace* as the column space $\mathcal{S}(\boldsymbol{B})$ of a matrix $\boldsymbol{B} \in \mathbb{R}^{p \times d}$, $d \leq p$, that satisfies

$$C_{\boldsymbol{X}} = C_{\boldsymbol{B}^\top \boldsymbol{X}}, \quad a.s. \tag{4.4}$$

If a matrix $\boldsymbol{B}$ satisfies (4.4), $\boldsymbol{X}$ and $\boldsymbol{B}^\top \boldsymbol{X}$ carry the same information about the dependence between the components of $\boldsymbol{Y}$. The smallest possible copula dimension reduction subspace is called the central copula subspace.

Definition 4.1 (Central copula subspace) Let $\mathcal{B}_{C_{\boldsymbol{X}}} = \{\boldsymbol{B} \in \mathbb{R}^{p \times p} : C_{\boldsymbol{X}} = C_{\boldsymbol{B}^\top \boldsymbol{X}}\ a.s.\}$. If $\mathcal{S}_{C_{\boldsymbol{X}}} = \bigcap_{\boldsymbol{B} \in \mathcal{B}_{C_{\boldsymbol{X}}}} \mathcal{S}(\boldsymbol{B})$ is a copula dimension reduction subspace, it is called the **central copula subspace**.

The following result states that the central subspace can be decomposed into the marginal central subspaces $\mathcal{S}_{Y_j|\boldsymbol{X}}$, $j = 1, \ldots, q$, and the central copula subspace $\mathcal{S}_{C_{\boldsymbol{X}}}$.

Theorem 4.1 *For the conditional copula $C_{\boldsymbol{X}}$ associated with the distribution of $\boldsymbol{Y} \mid \boldsymbol{X}$, for the central subspace it holds that $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} = \operatorname{span}\{\mathcal{S}_{Y_1|\boldsymbol{X}}, \dots, \mathcal{S}_{Y_q|\boldsymbol{X}}, \mathcal{S}_{C_{\boldsymbol{X}}}\}$.*

Proof Recall from the proof of Lemma 4.1 that $F_{j,\boldsymbol{X}}$ and $C_{\boldsymbol{X}}$ can be expressed in terms of $F_{\boldsymbol{Y}|\boldsymbol{X}}$. Hence, $\mathcal{S}_{Y_j|\boldsymbol{X}} \subset \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$, for $j = 1, \dots, q$, and $\mathcal{S}_{C_{\boldsymbol{X}}} \subset \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$, and, thus, $\operatorname{span}\{\mathcal{S}_{Y_1|\boldsymbol{X}}, \dots, \mathcal{S}_{Y_q|\boldsymbol{X}}, \mathcal{S}_{C_{\boldsymbol{X}}}\} \subset \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$. It also implies that any matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ with $\mathcal{S}(\boldsymbol{B}) = \mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ possesses submatrices $\boldsymbol{B}_j \in \mathbb{R}^{p\times d_j}$, $\boldsymbol{B}_C \in \mathbb{R}^{p\times d_C}$ satisfying $\mathcal{S}(\boldsymbol{B}_j) = \mathcal{S}_{Y_j|\boldsymbol{X}}$, $j = 1, \dots, q$, and $\mathcal{S}(\boldsymbol{B}_C) = \mathcal{S}_{C_{\boldsymbol{X}}}$. By Lemma 4.2, it follows that $\operatorname{span}\{\mathcal{S}_{Y_1|\boldsymbol{X}}, \dots, \mathcal{S}_{Y_q|\boldsymbol{X}}, \mathcal{S}_{C_{\boldsymbol{X}}}\}$ is a dimension reduction subspace for $F_{\boldsymbol{Y}|\boldsymbol{X}}$. And since $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}}$ is the intersection of all dimension reduction subspaces, it must hold $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} \subset \operatorname{span}\{\mathcal{S}_{Y_1|\boldsymbol{X}}, \dots, \mathcal{S}_{Y_q|\boldsymbol{X}}, \mathcal{S}_{C_{\boldsymbol{X}}}\}$. □

Theorem 4.1 can be seen as the central subspace version of Sklar's theorem. Similar to Lemma 4.1 and Lemma 4.2, the decomposition in Theorem 4.1 is not necessarily disjoint. But it is certainly possible that the central marginal or copula subspaces are of smaller dimension than the central subspace of the joint distribution.

Example 4.1 Consider a variant of the model from Example 3.2:

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{\beta}^\top \boldsymbol{X}) + \operatorname{diag}(\sigma_1, \dots, \sigma_q)\boldsymbol{R}(\boldsymbol{\gamma}^\top \boldsymbol{X})^{1/2}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{q\times q}),$$

where $\boldsymbol{R}$ is a map from $\mathbb{R}$ to the space of correlation matrices. The central subspace is again $\mathcal{S}_{\boldsymbol{Y}|\boldsymbol{X}} = \mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ and has dimension $d = 2$. For all $j = 1, \dots, q$, the marginal central subspaces is $\mathcal{S}_{Y_j|\boldsymbol{X}} = \mathcal{S}(\boldsymbol{\beta})$ with $d_j = 1$ (the marginal variances do not change with $\boldsymbol{X}$). The conditional copula $C_{\boldsymbol{X}}$ is a Gaussian copula with correlation matrix $\boldsymbol{R}(\boldsymbol{\gamma}^\top \boldsymbol{X})$. Hence, the central copula subspace is $\mathcal{S}_{C_{\boldsymbol{X}}} = \mathcal{S}(\boldsymbol{\gamma})$ and $d_C = 1$. □

The result in Theorem 4.1 is useful in practice because it facilitates inference. The popularity of copulas is largely due to the fact that they allow to separate inference of the marginal distributions and the copula. Similarly, inference of the central subspace can be separated into inference of the central marginal subspaces and the central copula subspace. On finite samples, smaller subspaces are usually easier to find. Further, when estimating the conditional margins and copula separately, we can condition on lower-dimensional covariates. This is particularly useful when the conditional margins are estimated nonparametrically, because it allows for faster rates of convergence.

Remark 4.1 The assumption $C_{\boldsymbol{X}} \equiv C$ or, equivalently, $C_{\boldsymbol{X}} \perp \boldsymbol{X}$ is known as the *simplifying assumption* in the copula literature (e.g., Gijbels et al. 2015b,a, 2017; Derumigny and Fermanian 2017; Portier and Segers 2018) and commonly used in vine copula models (e.g., Haff and Segers 2015; Spanhel et al. 2019; Portier and Segers 2018; Haff et al. 2010; Nagler and Czado 2016; Stoeber et al. 2013; Nagler 2024). It can be reformulated in terms of the central copula subspace: $\mathcal{S}_{C_{\boldsymbol{X}}} = \{\boldsymbol{0}\}$.

4.3 Properties of the central copula subspace

By definition of $C_{\boldsymbol{X}}$, the central copula subspace is the central subspace of the conditional random vector $\boldsymbol{U}_{\boldsymbol{X}} \mid \boldsymbol{X}$, i.e. $\mathcal{S}_{C_{\boldsymbol{X}}} = \mathcal{S}_{\boldsymbol{U}_{\boldsymbol{X}}|\boldsymbol{X}}$. Therefore, all remarks made in

Section 3 about the existence of the central subspace apply. We introduced the central copula subspace as a separate concept because it has distinct properties:

1. The marginal central subspaces only contain the zero element: one can easily verify that the variables $U_{j,\boldsymbol{X}}$ are marginally independent of $\boldsymbol{X}$ for $j = 1, \ldots, q$.
2. The random variables $U_{1,\boldsymbol{X}}, \ldots, U_{q,\boldsymbol{X}}$ are unobserved because the conditional margins $F_{Y_1|\boldsymbol{X}}$, ..., $F_{Y_q|\boldsymbol{X}}$ are unknown. This makes inference for the central copula subspace more difficult.

In view of Proposition 3.1, we can characterize the central copula subspace by a collection of central mean subspaces. In Section 3.3, we already mentioned two examples of such families which we shall discuss in more detail. The following result is an immediate consequence of Proposition 3.1.

Theorem 4.2 *Let*

$$\mathcal{G}_1 = \big\{g\colon \mathbb{R}^q \to \mathbb{R},\ \boldsymbol{z} \mapsto \mathbb{1}\big(\boldsymbol{z} \le \boldsymbol{u}\big),\ \boldsymbol{u} \in [0,1]^q\big\},$$
$$\mathcal{G}_2 = \left\{g\colon \mathbb{R}^q \to \mathbb{R},\ \boldsymbol{z} \mapsto \prod_{j=1}^{q} z_j^{k_j},\ (k_1, \ldots, k_q) \in \mathbb{N}^q\right\}.$$

Then, $\mathcal{S}_{C_{\boldsymbol{X}}} = \operatorname{span}\big\{S_{\mathrm{E}\{g(\boldsymbol{U}_{\boldsymbol{X}})|\boldsymbol{X}\}}, g \in \mathcal{G}_1\big\} = \operatorname{span}\big\{S_{\mathrm{E}\{g(\boldsymbol{U}_{\boldsymbol{X}})|\boldsymbol{X}\}}, g \in \mathcal{G}_2\big\}$.

Each conditional expectation $\mathrm{E}\{g(\boldsymbol{U}_{\boldsymbol{X}}) \mid \boldsymbol{X}\}$ in Theorem 4.2 summarizes a different aspect of $C_{\boldsymbol{X}}$. The families $\mathcal{G}_1$ and $\mathcal{G}_2$ are such that the collection of summaries is sufficient to characterize the conditional copula $C_{\boldsymbol{X}}$. The result is quite intuitive: For $\mathcal{G}_1$, the function $\boldsymbol{u} \mapsto \mathrm{E}\{\mathbb{1}(\boldsymbol{U}_{\boldsymbol{X}} \le \boldsymbol{u}) \mid \boldsymbol{X}\}$ is exactly the conditional copula $C_{\boldsymbol{X}}$. The collection $\mathcal{G}_2$ relates to the fact that the collection of all moments completely characterizes the distribution of a bounded random vector (Hausdorff 1921). In our context, we do not even require all moments, but only those where at least two exponents k_j are non-zero. This suffices because $\mathrm{E}(U_{j,\boldsymbol{X}}^k \mid \boldsymbol{X})$ is independent of $\boldsymbol{X}$ for all $k \ge 0$; see property (i) above.

The dependence summaries induced by $\mathcal{G}_1$ and $\mathcal{G}_2$ relate to popular conditional measures of association when $q = 2$ (see, Gijbels et al. 2011). The conditional Spearman's ρ can be expressed as $\rho_{\boldsymbol{X}} = 12\mathrm{E}\{U_{1,\boldsymbol{X}}U_{2,\boldsymbol{X}}|\boldsymbol{X}\} - 3$ which relates to the element of $\mathcal{G}_2$ where $k_1 = k_2 = 1$. The conditional Blomqvist's β can be expressed as $\beta_{\boldsymbol{X}} = 1 - 4\mathrm{E}\{\mathbb{1}(U_{1,\boldsymbol{X}} \le 0.5, U_{2,\boldsymbol{X}} \le 0.5)|\boldsymbol{X}\}$ and relates to the element of $\mathcal{G}_1$ with $u_1 = u_2 = 0.5$. In parametric models, these measures often suffice to characterize the whole dependence structure. In other cases, the dependence measure is the only quantity of interest. We discuss sufficient dimension reduction in this simplified setting in the following section.

5 Sufficient dimension reduction for conditional dependence measures

We now state the sufficient dimension reduction problem for conditional association measures and define central subspaces in this context. We discuss some popular measures and show that their central subspaces can be expressed in terms of central mean subspaces.

The simplest way to analyze dependence between two variables is to summarize it in a single number, a *measure of association* (see, e.g., Nelsen 2006, Chapter 5). Any dependence measure η satisfying Rényi's axiom F (invariance with respect to strictly increasing marginal transformation) (see, Rényi 1959) can be written as a functional T_η of the copula: $\eta = T_\eta(C)$ (see, Schweizer and Wolff 1981). The corresponding conditional measure of dependence $\eta_{\boldsymbol{X}}$ is defined as the same functional T_η applied to the conditional copula: $\eta_{\boldsymbol{X}} = T_\eta(C_{\boldsymbol{X}})$. The dimension reduction problem is now to find a matrix $\boldsymbol{B} \in R^{p\times d}$, with $d \leq p$, such that

$$\eta_{\boldsymbol{X}} = \eta_{\boldsymbol{B}^\top \boldsymbol{X}} \quad a.s. \tag{5.1}$$

We define an η *dimension reduction subspace* as the column space $\mathcal{S}(\boldsymbol{B})$ of any matrix satisfying (5.1). Our goal is to find the smallest of these spaces.

Definition 5.1 (Central η subspace) Let $\mathcal{B}_{\eta_{\boldsymbol{X}}} = \{\boldsymbol{B} \in \mathbb{R}^{p\times p} \colon \boldsymbol{\eta_X} = \eta_{\boldsymbol{B}^\top \boldsymbol{X}}\ a.s\}$. If $\mathcal{S}_{Y|\boldsymbol{X}} = \bigcap_{\boldsymbol{B}\in\mathcal{B}_{\eta_{\boldsymbol{X}}}} \mathcal{S}(\boldsymbol{B})$ is a η dimension reduction subspace, it is called the **central η subspace**.

In Section 4.3 we saw two examples of measures whose central η subspace is the central mean subspace related to a conditional expectation $\mathrm{E}\{g(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) \mid \boldsymbol{X}\}$. These examples belong to a more general class of association measures whose central subspaces can be expressed as central mean subspaces. Consequently, we can estimate the central η subspace using methods that are suitable for finding the central mean subspace.

Example 5.1 lists four popular dependence measures that are a linear functional of the conditional copula in the sense that there exists $g_\eta \colon [0,1]^2 \to \mathbb{R}$ such that $\eta_{\boldsymbol{X}} = \int g_\eta(u_1,u_2) dC_{\boldsymbol{X}}(u_1,u_2) = \mathrm{E}\{g_\eta(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) \mid \boldsymbol{X}\}$. Hence, $\mathcal{S}_{\eta_{\boldsymbol{X}}} = \mathcal{S}_{\mathrm{E}\{g_\eta(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}|\boldsymbol{X})\}}$.

Example 5.1

- Spearman's ρ: $g_\rho(u_1,u_2) = 12(u_1 - 0.5)(u_2 - 0.5)$,
- Blomqvist's β: $g_\beta = 1 - 4\mathbb{1}(u_1 \leq 0.5, u_2 \leq 0.5)$,
- Gini's γ: $g_\gamma = 2\big(|u_1 + u_2 - 1| - |u_1 - u_2|\big)$,
- van der Waerden's coefficient: $g_\omega = \Phi^{-1}(u_1)\Phi^{-1}(u_2)$, where Φ^{-1} is the standard normal quantile function. □

All measures above belong to the family of *concordance measures* (Nelsen 2006). The conditional Kendall's τ is another popular concordance measure that is not of the form $\mathrm{E}\{g_\eta(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) \mid \boldsymbol{X}\}$. We can still relate it to a central mean subspace, but in a more complicated fashion. The conditional Kendall's τ can be expressed as $\tau_{\boldsymbol{X}} = 4\int C_{\boldsymbol{X}}(u_1,u_2)dC_{\boldsymbol{X}}(u_1,u_2)-1$, see Gijbels et al. (2011). Let $(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) \sim C_{\boldsymbol{X}}$, $(\bar{U}_{1,\boldsymbol{X}}, \bar{U}_{2,\boldsymbol{X}}) \sim C_{\boldsymbol{X}}$ be two vectors such that $(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) \perp (\bar{U}_{1,\boldsymbol{X}}, \bar{U}_{2,\boldsymbol{X}}) \mid \boldsymbol{X}$. Then,

$$\int C_{\boldsymbol{X}}(u_1,u_2)dC_{\boldsymbol{X}}(u_1,u_2) = P\big(U_{1,\boldsymbol{X}} \le \bar{U}_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}} \le \bar{U}_{2,\boldsymbol{X}} \mid \boldsymbol{X}\big),$$

see Nelsen (2006, p. 160) for a derivation in the unconditional case. Hence, $\tau_{\boldsymbol{X}} = \tau_{\boldsymbol{B}^\top \boldsymbol{X}}$ a.s., if and only if, a.s.,

$$\mathrm{E}\big\{g_\tau\big(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}, \bar{U}_{1,\boldsymbol{X}}, \bar{U}_{2,\boldsymbol{X}}\big) \mid \boldsymbol{X}\big\} = \mathrm{E}\big\{g_\tau\big(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}, \bar{U}_{1,\boldsymbol{X}}, \bar{U}_{2,\boldsymbol{X}}\big) \mid \boldsymbol{B}^\top \boldsymbol{X}\big\}, \tag{5.2}$$

where $g_\tau(u_1,u_2,\bar{u}_1,\bar{u}_2) = 4\mathbb{1}(u_1 \le \bar{u}_2, u_2 \le \bar{u}_2) - 1$. Although the expressions in (5.2) are more difficult to estimate due to the presence of conditionally independent copies, they are still conditional expectations. Hence, the central τ subspace is also a central mean subspace:

$$\mathcal{S}_{\tau_{\boldsymbol{X}}} = \mathcal{S}_{\mathrm{E}\{g_\tau(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}, \bar{U}_{1,\boldsymbol{X}}, \bar{U}_{2,\boldsymbol{X}} \mid \boldsymbol{X})\}}.$$

6 Differences and similarities between central subspaces for conditional dependence

Since the conditional measures of association $\eta_{\boldsymbol{X}}$ discussed in the previous section are functionals of the conditional copula, it generally holds that $\mathcal{S}_{\eta_{\boldsymbol{X}}} \subseteq \mathcal{S}_{C_{\boldsymbol{X}}}$. This implies that the dimension can possibly be reduced further when a measure of association is of concern. Another question concerns the relationship between central η subspaces for different choices of η. Universal statements of this kind seem out of reach. However, all subspaces do coincide in many situations of practical interest. Let us substantiate this by an example.

Example 6.1 Consider a parametric family of copulas $\mathcal{C}^\theta = \{C(\cdot,\cdot;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}\}$ and let $h \colon \Omega_{\boldsymbol{X}} \to \Theta$ be unknown, with $\Omega_{\boldsymbol{X}}$ the support of $\boldsymbol{X}$ ($\Omega_{\boldsymbol{X}} \subseteq \mathbb{R}^q$). Then, a general semiparametric model for the conditional copula is

$$\big(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}\big) | \boldsymbol{X} = \boldsymbol{x} \sim C\big\{\cdot,\cdot; h(\boldsymbol{x})\big\}, \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^p. \tag{6.1}$$

For most of the popular parametric families there is a one-to-one relationship between the dependence parameter θ and the measures of concordance $T_\eta(C^\theta)$. Examples include the Gaussian, Student t (with fixed degrees of freedom), Farlie-Gumbel-Morgenstern (FGM), Clayton, Gumbel, Joe, and Frank copula families.[1] When the relationship between θ and $T_\eta(C^\theta)$ is one-to-one, the covariate $\boldsymbol{X}$ always affects the copula and η at the same time. Similarly, for any fixed matrix $\boldsymbol{B} \in \mathbb{R}^{p\times d}$, a change in $\eta_{\boldsymbol{B}^\top \boldsymbol{X}}$ always leads to a change in $C_{\boldsymbol{B}^\top \boldsymbol{X}}$ and the other way around. Hence,

$$\mathcal{S}_{C_{\boldsymbol{X}}} = \mathcal{S}_{\rho_{\boldsymbol{X}}} = \mathcal{S}_{\beta_{\boldsymbol{X}}} = \mathcal{S}_{\gamma_{\boldsymbol{X}}} = \mathcal{S}_{\omega_{\boldsymbol{X}}} = \mathcal{S}_{\tau_{\boldsymbol{X}}}.$$

[1] Although analytical formulas for the relationship rarely exist, one can check numerically that the relation is in fact one-to-one.

The models considered by Patton (2001), Acar et al. (2011), Abegaz et al. (2012), Vatter and Chavez-Demoulin (2015), and Fermanian and Lopez (2018) are all of this type. $\square$

On the other hand, we emphasize that this is not true in general. Let us give an obvious counterexample where the central copula subspace is different from central η subspaces.

Example 6.2 Consider model (6.1) for the Student t copula where the association parameter is fixed, and θ refers to the degrees of freedom parameter. The degrees of freedom has no influence on Kendall's τ, i.e., $\tau_{\boldsymbol{X}}$ is constant. Then it holds trivially that $\tau_{\boldsymbol{X}} = \tau_{\boldsymbol{B}^\top \boldsymbol{X}}$ for $\boldsymbol{B} = \mathbf{0}$ which implies $\mathcal{S}_{\tau_{\boldsymbol{X}}} = \{\mathbf{0}\}$. However, $\mathcal{S}_{C_{\boldsymbol{X}}} \neq \{\mathbf{0}\}$, because $C_{\boldsymbol{X}}$ does depend on the degrees of freedom parameter and a change in $h(\boldsymbol{X})$ will have an effect on $C_{\boldsymbol{X}}$. $\square$

The next example demonstrates that it is also possible that central η subspaces are different for different association measures.

Example 6.3 Consider two parametric copula families

$$\mathcal{C}_1^{\theta_1} = \{C_1(\cdot,\cdot;\theta_1)\colon \theta_1 \in \Theta_1 \subseteq \mathbb{R}\}, \quad \mathcal{C}_2^{\theta_2} = \{C_2(\cdot,\cdot;\theta_2)\colon \theta_2 \in \Theta_2 \subseteq \mathbb{R}\}.$$

For $\theta_1 \in \Theta_1, \theta_2 \in \Theta_2$, and a mixing parameter $\pi \in [0,1]$, we define the mixture of the two copulas $C(u_1,u_2;\theta_1,\theta_2,\pi) = \pi C_1(u_1,u_2;\theta_1) + (1-\pi)C_2(u_1,u_2;\theta_2)$. Let $p = p_1 + p_2$, $h_1\colon \mathbb{R}^{p_1} \to \Theta_1$, $h_2\colon \mathbb{R}^{p_2} \to \Theta_2$, $\pi\colon \mathbb{R}^{p_2} \to [0,1]$. We consider the following conditional copula model: for all $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathbb{R}^p$,

$$\big(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}\big) | \boldsymbol{X} = \boldsymbol{x} \sim C\big\{\cdot,\cdot; h_1(\boldsymbol{x}^{(1)}), h_2(\boldsymbol{x}^{(1)}), \pi(\boldsymbol{x}^{(2)})\big\}.$$

In this model, the first component of the covariate vector, $\boldsymbol{x}^{(1)}$, affects the strength of dependence in both parts of the mixture. The second component, $\boldsymbol{x}^{(2)}$, controls the weighting between the two parts.

We have $\mathcal{S}_{C_{\boldsymbol{X}}} = \operatorname{span}(\boldsymbol{e}^{(1)}, \boldsymbol{e}^{(2)})$, where $\boldsymbol{e}^{(2)}$ is a vector with 1 in the first p_1 components and all remaining components equal 0, and $\boldsymbol{e}^{(2)} = \mathbf{1} - \boldsymbol{e}^{(1)}$. Assume further that h_1 and h_2 are such that for all $\boldsymbol{x}^{(1)} \in \mathbb{R}^{p_1}$,

$$\int g_\rho(u_1,u_2) dC_1\{u_1,u_2; h_1(\boldsymbol{x}^{(1)})\} = \int g_\rho(u_1,u_2) dC_2\{u_1,u_2\cdot; h_2(\boldsymbol{x}^{(1)})\}. \tag{6.2}$$

Since $\pi(\boldsymbol{x}^{(2)}) + (1 - \pi(\boldsymbol{x}^{(2)})) = 1$, the Spearman's ρ of C is not affected by $\boldsymbol{x}^{(2)}$. Hence, $\mathcal{S}_{\rho_{\boldsymbol{X}}} = \operatorname{span}(\boldsymbol{e}_1)$. However, it is generally not the case that

$$\int g_\beta(u_1,u_2) dC_1\{u_1,u_2; h_1(\boldsymbol{x}^{(1)})\} = \int g_\beta(u_1,u_2) dC_2\{u_1,u_2\cdot; h_2(\boldsymbol{x}^{(1)})\} \tag{6.3}$$

holds at the same time. If it does not, we have $\mathcal{S}_{\rho_{\boldsymbol{X}}} \neq \mathcal{S}_{\beta_{\boldsymbol{X}}}$. This is easy to check when analytical expressions for the relationship between the measures and the copula parameter are available. For example, it is not possible that (6.2) and (6.3) hold at the same time when $\mathcal{C}_1^{\theta_1}$ is the Gaussian family and $\mathcal{C}_2^{\theta_2}$ is the FGM family. $\square$

We conclude that the various central subspaces may differ in conditional dependence models. But in most situations of practical interest, they coincide.

7 Estimation of central subspaces in conditional dependence models

In Sections 4 and 5 we have seen that sufficient dimension reduction problems in conditional dependence models can be transformed to equivalent dimension reduction problems in a classical (mean) regression setting. A large number of methods for the estimation of central (mean) subspaces have been proposed (see, e.g., Ma and Zhu 2013). We focus on a particular instance of nonparametric estimators: the *outer product of gradients (OPG)* method which was introduced by Xia et al. (2002) and is based on an idea of Härdle and Stoker (1989). The main motivation for this choice is the simplicity and generality of the OPG method. It is often used as a starting point for other techniques that need to be initialized with a consistent estimate of the central subspace (e.g., Xia et al. 2002; Xia 2007; Luo et al. 2014). Of course, all other methods, e.g., the ones from Ma and Zhu (2012) or Huang and Chiang (2017), are similarly applicable in the context of conditional dependence models.

7.1 The OPG matrix and central subspaces

Let $\mathcal{G} \subset \{g\colon \mathbb{R}^q \to \mathbb{R}\}$ be a set of functions and we want to identify the space $\mathcal{S}_{\mathcal{G}} = \operatorname{span}\{\mathcal{S}_{\mathrm{E}\{g(\boldsymbol{Y})|\boldsymbol{X}\}}\colon g \in \mathcal{G}\}$. For central mean subspaces, $\mathcal{G}$ is a singleton. A larger, but finite number of suitably chosen functions suffice for central subspaces, see Proposition 3.1 and the following remarks. In the following, we thus assume $|\mathcal{G}| < \infty$ for simplicity (this can be relaxed by randomizing over elements in $\mathcal{G}$, see, Yin and Li 2011). Define

$$m_g(\boldsymbol{x}) = \mathrm{E}\{g(\boldsymbol{Y}) \mid \boldsymbol{X} = \boldsymbol{x}\}, \qquad m_g^{\boldsymbol{B}}(\boldsymbol{x}) = \mathrm{E}\{g(\boldsymbol{Y}) \mid \boldsymbol{B}^\top \boldsymbol{X} = \boldsymbol{B}^\top \boldsymbol{x}\},$$
$$\boldsymbol{\Delta}_{\mathcal{G}} = \sum_{g\in\mathcal{G}} \mathrm{E}\big\{\nabla m_g(\boldsymbol{X})\nabla^\top m_g(\boldsymbol{X})\mathbb{1}_{\boldsymbol{X}\in\mathcal{D}_{\boldsymbol{X}}}\big\},$$

where $\mathcal{D}_{\boldsymbol{X}} \subseteq \mathbb{R}^p$ is some set. The OPG method is based on the following observation.

Lemma 7.1 *Suppose* $\boldsymbol{B}_0 \in \mathbb{R}^{p\times d}$ *is a matrix such that* $\operatorname{span}(\boldsymbol{B}_0) = \mathcal{S}_{\mathcal{G}}$ *and*

$$\boldsymbol{\Delta}_{\mathcal{G}}^{\boldsymbol{B}_0} = \sum_{g\in\mathcal{G}} \mathrm{E}\{\nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})\nabla^\top m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})\mathbb{1}_{\boldsymbol{X}\in\mathcal{D}_{\boldsymbol{X}}}\}$$

is of full rank. Let $\boldsymbol{\Delta}_{\mathcal{G}} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$ *be the eigen-decomposition of* $\boldsymbol{\Delta}_{\mathcal{G}}$*. Then* $\mathcal{S}_{\mathcal{G}} = \operatorname{span}(\boldsymbol{V}\boldsymbol{\Lambda}^{1/2})$.

Proof Recall that since $\Delta_{\mathcal{G}}$ is symmetric, $\boldsymbol{V} = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$ is orthogonal and $\boldsymbol{\Lambda}$ is diagonal containing the eigenvalues on the diagonal. Assume without loss of generality that they are ordered such that $\lambda_1 \geq \ldots \geq \lambda_d > 0$ and $\lambda_{d+1} = \ldots = \lambda_p = 0$. Observe that $\operatorname{span}(\boldsymbol{V}\boldsymbol{\Lambda}^{1/2}) = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_d)$ and recall that we have the orthogonal decomposition $\mathbb{R}^p = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_d) + \operatorname{span}(\boldsymbol{v}_{d+1}, \dots, \boldsymbol{v}_p)$. Since by assumption $m_g(\boldsymbol{X}) = m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})$, it holds that $\nabla m_g(\boldsymbol{X}) = \nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})\boldsymbol{B}_0^\top$, for all $g \in \mathcal{G}$. We thus have

$$\boldsymbol{\Delta}_{\mathcal{G}} = \sum_{g\in\mathcal{G}} \boldsymbol{B}_0\mathrm{E}\big\{\nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})^\top \nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X})\big\}\boldsymbol{B}_0^\top = \boldsymbol{B}_0\boldsymbol{\Delta}_{\mathcal{G}}^{\boldsymbol{B}_0}\boldsymbol{B}_0^\top.$$

Let $\boldsymbol{v} \neq \boldsymbol{0}$ be some vector. Then because $\boldsymbol{\Delta}_{\mathcal{G}}^{\boldsymbol{B}_0}$ has full rank, we have

$$\boldsymbol{v} \in \operatorname{span}(\boldsymbol{v}_{d+1}, \dots, \boldsymbol{v}_p) \quad \Leftrightarrow \quad \boldsymbol{v}^\top \boldsymbol{\Delta}_{\mathcal{G}} \boldsymbol{v} = 0 \quad \Leftrightarrow \quad \boldsymbol{v}^\top \boldsymbol{B}_0 \boldsymbol{\Delta}_{\mathcal{G}}^{\boldsymbol{B}_0} \boldsymbol{B}_0^\top \boldsymbol{v} = 0 \quad \Leftrightarrow \quad \boldsymbol{v} \perp \mathcal{S}_{\mathcal{G}}.$$

The orthogonal complement of $\operatorname{span}(\boldsymbol{v}_{d+1}, \dots, \boldsymbol{v}_p)$ is $\operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_d)$, so negating the statements above gives

$$\boldsymbol{v} \in \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_d) \quad \Leftrightarrow \quad \boldsymbol{v} \in \mathcal{S}_{\mathcal{G}}. \qquad \square$$

The assumption that $\boldsymbol{\Delta}_{\mathcal{G}}^{\boldsymbol{B}_0}$ is of full rank can be deciphered as follows: (i) $\boldsymbol{B}_0$ does not contain any redundant columns, (ii) the restriction of m_g to the set $\mathcal{D}_{\boldsymbol{X}}$ covers all relevant directions of variation. The lemma implies that the central subspace is spanned by the eigenvectors of the OPG matrix corresponding to non-zero eigenvalues. In particular, if $\dim(\mathcal{S}_{\mathcal{G}}) = d$, the first d eigenvectors of $\boldsymbol{\Delta}_{\mathcal{G}}$ span $\mathcal{S}_{\mathcal{G}}$.

7.2 An adaptive OPG method for estimating central subspaces

We introduce an adaptive variant of the method, similar to the ones proposed by Xia (2007) and Yin and Li (2011). Because our main consistency result may be of independent interest, we present the method and its convergence properties in a general setting first. Estimation of central dependence subspaces is discussed in the following section.

Let $\boldsymbol{X}_{i,\boldsymbol{x}} = \boldsymbol{X}_i - \boldsymbol{x}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times r}$, $1 \leq r \leq p$. To estimate the gradients $\nabla m_g(\boldsymbol{x})$, define the local-linear regression estimator

$$\begin{pmatrix} \widehat{m}_g(\boldsymbol{x}) \\ \widehat{\nabla} m_g(\boldsymbol{x}) \end{pmatrix} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left\{ g(\boldsymbol{Y}_i) - \boldsymbol{\beta}^\top \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} \right\}^2 K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}), \tag{7.1}$$

where $K_h(\boldsymbol{z}) = h^{-r} \prod_{k=1}^r K(z_k/h)$ for a univariate kernel function K. By $K(\boldsymbol{z})$ we denote the multivariate product kernel, i.e. $K(\boldsymbol{z}) = \prod_{k=1}^r K(z_k)$.

The corresponding estimator for the OPG matrix is

$$\widehat{\boldsymbol{\Delta}}_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \widehat{\nabla} m_g(\boldsymbol{X}_i) \widehat{\nabla}^\top m_g(\boldsymbol{X}_i) \mathbb{1}(\boldsymbol{X}_i \in \mathcal{D}_{\boldsymbol{X}}), \tag{7.2}$$

for some compact set $\mathcal{D}_{\boldsymbol{X}} \subset \mathbb{R}^p$ introduced to stabilize the estimator. An estimate for the central subspace $\mathcal{S}_{\mathcal{G}}$ can be obtained by computing the eigenvectors of $\widehat{\boldsymbol{\Delta}}_{\mathcal{G}}$ corresponding to the d largest eigenvalues. However, because the gradients are estimated nonparametrically, this estimate $\mathcal{S}_{\mathcal{G}}$ suffers from the curse of dimensionality. If $p \gg d$, the convergence will be extremely slow.

To overcome this, we consider the following adaptive procedure. Supposing for simplicity that $d = \dim(\mathcal{S}_{\mathcal{G}})$ is known, set $\widehat{\boldsymbol{B}}^{(0)} = \boldsymbol{I}_{p \times p}$, $h_0 \sim (\ln n/n)^{1/(6+p)}$ and fix some $\rho \in (0,1)$ and $h_\infty \in (0, \infty)$. For $t = 0, 1, 2, \dots$:

1. Set $h_t = \max\{\rho h_{t-1}, h_\infty\}$.
2. Compute $\widehat{\nabla} m_g(\boldsymbol{x})$ in (7.1) with $\boldsymbol{B} = \widehat{\boldsymbol{B}}^{(t)}$ and $h = h_t$ and compute $\widehat{\boldsymbol{\Delta}}_{\mathcal{G}}$ as in (7.2).

3. Define $\widehat{\boldsymbol{B}}^{(t+1)} = \widehat{\boldsymbol{V}}\mathrm{diag}\{s(\widehat{\boldsymbol{\Lambda}})\}$ where $\widehat{\boldsymbol{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{V}}^\top$ is the eigen-decomposition of $\widehat{\boldsymbol{\Delta}}_{\mathcal{G}}$, and $s\colon \mathbb{R}^{p\times p} \to \mathbb{R}^p$ satisfies $s(\boldsymbol{\Lambda})_j = 0$ if $\Lambda_{j,j} = 0$ for $j > d$.

The function s is introduced to stabilize the convergence of the algorithm. Practical recommendations for the hyperparameters are given in Section 8.1.2.

Denote by $\widehat{\boldsymbol{B}}^{(\infty)}$ the estimate obtained after convergence of this adaptive procedure.

Assumptions

A1 The kernel K is a Lipschitz, symmetric probability density with support $[-1, 1]$.

A2 The density $f_{\boldsymbol{B}_0^\top \boldsymbol{X}}$ has two bounded continuous derivatives on $\mathcal{D}_{\boldsymbol{X}}$, and $\inf_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{X}}} f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}) > 0$.

A3 The matrix valued function

$$\boldsymbol{M}(\boldsymbol{x}) = \mathrm{E}\left\{ \begin{pmatrix} 1 & \boldsymbol{X}_{\boldsymbol{x}}^\top \\ \boldsymbol{X}_{\boldsymbol{x}} & \boldsymbol{X}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}}^\top \end{pmatrix} \middle| \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x} \right\},$$

with $\boldsymbol{X}_{\boldsymbol{x}} = \boldsymbol{X} - \boldsymbol{x}$, has eigenvalues bounded away from zero, and is twice continuously differentiable, uniformly on $\mathcal{D}_{\boldsymbol{X}}$.

A4 The functions $m_g, g \in \mathcal{G}$, have four continuous derivatives on $\mathcal{D}_{\boldsymbol{X}}$.

A5 The eigenvalues of $\boldsymbol{\Delta}_{\mathcal{G}} = \boldsymbol{V}_0\boldsymbol{\Lambda}_0\boldsymbol{V}_0^\top$ satisfy $\lambda_1 > \ldots > \lambda_d > \lambda_{d+1} = \ldots = \lambda_p = 0$, and the first d columns of $\boldsymbol{B}_0 = \boldsymbol{V}_0\mathrm{diag}\{s(\boldsymbol{\Lambda}_0)\}$ span $\mathcal{S}_{\mathcal{G}}$.

To avoid ambiguity, we shall always assume that the eigenvectors of $\widehat{\boldsymbol{B}}^{(\infty)}$ and of $\boldsymbol{B}_0 = \boldsymbol{V}_0\mathrm{diag}\{s(\boldsymbol{\Lambda}_0)\}$ are oriented such that their first component is non-negative. Let $\|\cdot\|$ denote the operator norm for matrices. The following result is proven in Section A.3.

Theorem 7.1 *Let $h_\infty \to 0$, $nh_\infty^d/\ln n \to \infty$, as $n \to \infty$, and define $r_{n,d,\infty} = (nh_\infty^d/\ln n)^{-1/2}$. Under A1–A5, $\|\widehat{\boldsymbol{B}}^{(\infty)} - \boldsymbol{B}_0\| = O_p(n^{-1/2} + h_\infty^4 + r_{n,d,\infty}^2)$.*

The convergence rate in Theorem 7.1 does not depend on the number of covariates p, but the reduced dimension d. If $d \leq 3$, we can choose h_∞ such that $\|\widehat{\boldsymbol{B}}^{(\infty)} - \boldsymbol{B}_0\| = O_p(n^{-1/2})$; for example $h_\infty \sim n^{-1/8}/\ln n$. A similar result was proven by Xia (2007) (who used a slightly different adaptation strategy), but only for a specific family $\mathcal{G}$ consisting of smoothing kernels.

7.3 Estimation of central dependence subspaces

We now turn to estimation of the central dependence subspaces. We consider a finite set of functions $\mathcal{G}$ such that $\mathcal{S}_{\mathcal{G}} = \mathrm{span}\{\mathcal{S}_{\mathrm{E}\{g(\boldsymbol{U}_{\boldsymbol{X}})|\boldsymbol{X}\}}\}$. For central η subspaces, $\mathcal{G}$ is a singleton. A larger, but finite number of suitably chosen functions suffice for central copula subspaces, see Theorem 4.2 and the following remarks. In most practically relevant cases, the central copula subspace can be identified by a single function; see Section 6.

Assume for now that the conditional margins are known. For $\ell = 1, \dots, q$, $i = 1, \dots, n$, set $U_{\ell,\boldsymbol{X}_i} = F_{Y_\ell|\boldsymbol{X}}(Y_{i,\ell} \mid \boldsymbol{X}_i)$. Define

$$\begin{pmatrix} \widehat{m}_g(\boldsymbol{x}) \\ \widehat{\nabla} m_g(\boldsymbol{x}) \end{pmatrix} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left\{ g(\boldsymbol{U}_{\boldsymbol{X}_i}) - \boldsymbol{\beta}^\top \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} \right\}^2 K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}),$$

$$\widehat{\boldsymbol{\Delta}}_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \widehat{\nabla} m_g(\boldsymbol{X}_i) \widehat{\nabla}^\top m_g(\boldsymbol{X}_i) \mathbb{1}(\boldsymbol{X}_i \in \mathcal{D}_{\boldsymbol{X}}).$$

Setting $\boldsymbol{Y}_i = \boldsymbol{U}_{\boldsymbol{X}_i}$ in Theorem 7.1, we get

$$\|\widehat{\boldsymbol{B}}^{(\infty)} - \boldsymbol{B}_0\| = O_p(n^{-1/2} + h_\infty^4 + r_{n,d,\infty}^2).$$

The right-hand side is of order $O_p(n^{-1/2})$ if $d \leq 3$ and h_∞ appropriately chosen, see the comments after Theorem 7.1.

The situation is more complicated when margins have to be estimated, because the estimation error propagates. When a correctly specified parametric model is estimated, we can expect these errors to contribute at most a term of order $O_p(n^{-1/2})$. In the non-parametric case, we can use the adaptive OPG method to estimate the central marginal subspaces. The convergence rate of resulting marginal estimators would be $O_p(h_\ell^2 + (nh^{d_\ell}/\ln n)^{-1/2})$, where h_ℓ and d_ℓ are the bandwidth and dimension for estimating the ℓ-th marginal distribution. This is of larger order than $O_p(n^{-1/2})$ no matter the choice of h_ℓ. Some preliminary considerations suggest that the error may still be negligible when higher-order properties of the marginal estimators are taken into account. A complete analysis would involve fairly complicated third-order U-statistics. We shall not pursue this any further and refer to the simulations in the next section for an empirical evaluation.

8 Simulation study

In the following section, we present a simulation study to illustrate the performance of the proposed estimators. We compare the non-adaptive OPG estimator with a single iteration, the adaptive OPG estimator, and a parametric estimator. The simulation study is based on the following setup.

8.1 Setup

The goal of our study is to identify situations where sufficient dimension can be expected to work. To achieve this, we use a relatively simple simulation model for the covariates and conditional dependence and investigate the effect of its hyperparameters.

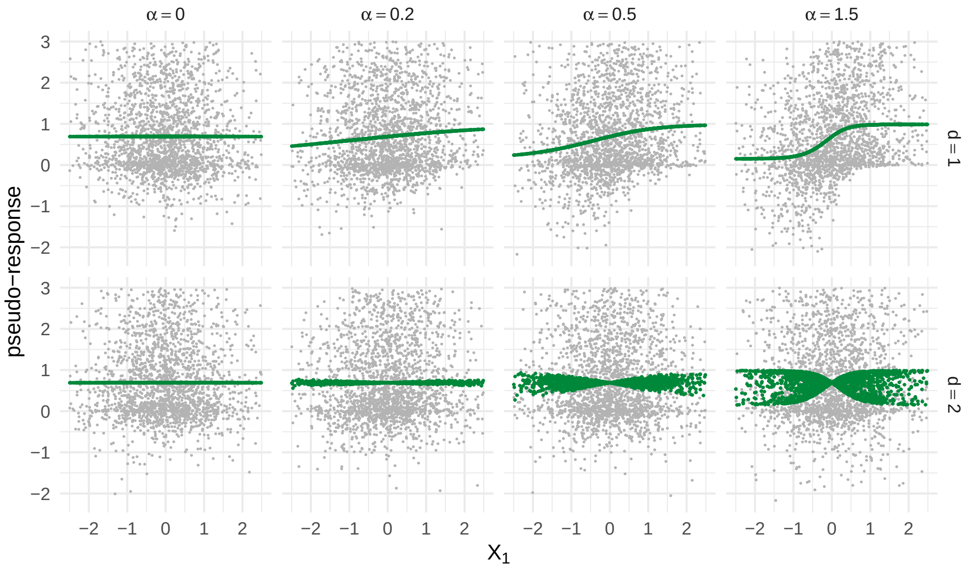

Fig. 1: Scatter plot of the active covariate X_1 against the pseudo-response for Spearman's ρ under known margins (gray) and the conditional Spearman's ρ (green).

8.1.1 Simulation model

In particular, we draw $\boldsymbol{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma = \frac{1}{2}\boldsymbol{I}_{p\times p} + \frac{1}{2}\boldsymbol{1}_{p\times p}$ (where $\boldsymbol{1}_{p\times p}$ denotes the $p \times p$ matrix with all elements equal to one) and define the conditional distribution of $(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}})$ by a parametric copula (Gaussian and Clayton) with conditional Kendall's τ given $\boldsymbol{X} = \boldsymbol{x}$ given as

$$\tau(\boldsymbol{x}) = \frac{1}{2} + \frac{2}{5}\prod_{j=1}^{d} \tanh(\alpha x_j).$$

Here, $\alpha \in \mathbb{R}$ is a parameter that controls the strength of the signal, and $\boldsymbol{e}_j$ is the jth vector of the standard basis in $\mathbb{R}^p$. The central dependence subspaces all coincide and equal $S_{C_{\mathbf{X}}} = \operatorname{span}(\boldsymbol{e}_1, \dots, \boldsymbol{e}_d)$. The results were found to be relatively robust to changes in the copula family, covariate distribution, and correlation among the covariates in preliminary experiments. Lastly, the responses Y_1, Y_2 are set to

$$Y_1 = \frac{1}{5}X_4^2 + \frac{1}{5}X_5^2 + \Phi^{-1}(U_{1,\boldsymbol{X}}), \qquad Y_2 = -X_2 - \frac{1}{5}X_4^2 + \Phi^{-1}(U_{2,\boldsymbol{X}}).$$

The central marginal subspaces are $\mathcal{S}_{Y_1|\boldsymbol{X}} = \operatorname{span}(\boldsymbol{e}_4, \boldsymbol{e}_5)$ and $\mathcal{S}_{Y_2|\boldsymbol{X}} = \operatorname{span}(\boldsymbol{e}_2, \boldsymbol{e}_4)$. Both are different from the central copula subspace.

To get a sense of the difficulty of the problem, we plot pseudo-response for Spearman's ρ, i.e., $g(U_{1,\boldsymbol{X}}, U_{2,\boldsymbol{X}}) = 12(U_{1,\boldsymbol{X}} - \frac{1}{2})(U_{2,\boldsymbol{X}} - \frac{1}{2})$ as a function of X_1 in Figure 1 (gray points). The conditional Spearman's rho is shown as green points. The left panel ($\alpha = 0$) correspond to no effect; this is how the data looks in directions orthogonal to the central copula subspace. The other panels show increasing an effect of the covariate on the conditional dependence from left to right. In all upper panels, we observe that the signal-to-noise ratio is relatively low — even for the strongest signal ($\alpha = 1.5$), where $\rho(X_1)$ changes rapidly from almost 0 to 1 in the center of the covariate distribution. In the lower panel ($d = 2$), the conditional Spearman's ρ is also modulated by X_2 which further obscures the relationship between X_1 and the pseudo-response. The signal-to-noise ratio is further decreased when the margins are estimated. This is in stark contrast to the well-behaved settings that methods for estimating the central mean subspace are usually evaluated in, where the signal-to-noise ratio is often very high. This is an inherent difficulty of estimating conditional dependence, because a single observation carries much less information.

8.1.2 Estimators

As explained in Section 6, the central copula subspace and central η subspaces coincide in most cases of practical interest. We therefore focus on the performance of three methods for estimating the central η subspace:

- `par`: parametric estimator in a correctly specified model; numerical optimization starting at the true value of $\boldsymbol{B}$.
- `OPG1`: The OPG estimator for the central η subspace after a single iteration.
- `OPGA`: The adaptive OPG estimator for the central η subspace.

The parametric estimator is included as a (practically infeasible) baseline to assess the limits of what can be achieved in a given setting. The OPG estimator is implemented with the hyperparameters

$$h_0 = n^{-1/(6+p)}, \quad \rho = n^{-1/(12+2p)}, \quad h_\infty = n^{-1/(4+d)},$$

in line with the recommendations of Xia (2007). Additionally, we use the stabilizing transformation

$$s(\boldsymbol{\Lambda}) = (s_1, \ldots, s_1, s_{d+1}, \ldots, s_p), \quad \text{where} \quad s_j = \Lambda_{jj}\left(\frac{1}{2} + \frac{1}{2\sum_{j=1}^p \Lambda_{j,j}}\right),$$

which dramatically improved the convergence behavior, especially in small-sample and weak signal settings.

In addition, we consider three settings for the conditional margins $F_{Y_j|\boldsymbol{X}}$:

- `known`: The conditional margins are known.
- `parametric`: The conditional margins estimated by maximum likelihood in the correctly specified parametric model.
- `nonparametric`: The conditional margins estimated by first estimating the marginal central mean subspace using OPG and then using a local linear

estimator with kernel smoothing for Y_j, i.e.,

$$\widehat{F}_{Y_j|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \arg\min_\alpha \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{ \int_{\infty}^{y} K_b(s - Y_{i,j}) ds - \alpha - \boldsymbol{\beta}^\top \boldsymbol{X}_{i,\boldsymbol{x}} \right\}^2 K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}).$$

The bandwidths for the OPG part are as above and the bandwidths for the conditional marginal distributions are set to $h = n^{-1/4}$ and $b = n^{-1/2}$. This strong under-smoothing is suggested by an informal analysis of the asymptotic properties under estimated margins and seen to work well in practice.

8.1.3 Performance measure

Performance is measured by the metric

$$d(\widehat{\boldsymbol{B}}, \boldsymbol{B}_0) = \|\widehat{\boldsymbol{B}}\widehat{\boldsymbol{B}}^\top - \boldsymbol{B}_0\boldsymbol{B}_0^\top\|_2,$$

where $\boldsymbol{B}_0 = (\boldsymbol{e}_1, \dots, \boldsymbol{e}_d)$ and $\widehat{\boldsymbol{B}}$ is an orthonormal basis of the estimated central η-subspace. The metric takes into account the non-identifiability of the basis matrices by comparing projectors onto the corresponding subspaces.

8.1.4 Scenarios

We consider the following scenarios:

- *Sample size*: $n = 150, 400, 1000, 2500$.
- *Covariate dimension*: $p = 5, 10, 20$.
- *Subspace dimension*: $d = 1, 2$.
- *Signal strength*: $\alpha = 0.2, 0.5, 1.5$.

Results are based on 100 replications for each scenario.

8.2 Results

The results for estimating the central ρ subspace under conditionally Gaussian dependence are shown in Figure 2. In each panel, the sample size is shown on the x-axis and the average error in estimating the central copula subspace is shown on the y-axis. Both are on log-scale so that a linear trend with slope $-1/2$ corresponds to $\sqrt{n}$-convergence. The different panels correspond to different signal strengths (α), covariate dimensions (p), and subspace dimensions (d). A few observations can be made:

- As expected, the error is increasing in d and p and decreasing in α.
- The adaptive OPG method always improves on the non-adaptive OPG method.
- Some low-signal settings are so difficult that not even the parametric estimator in the correctly specified model gives reasonable results for small sample sizes or larger d and p. (The better performance of the parametric estimator for $p = 20$

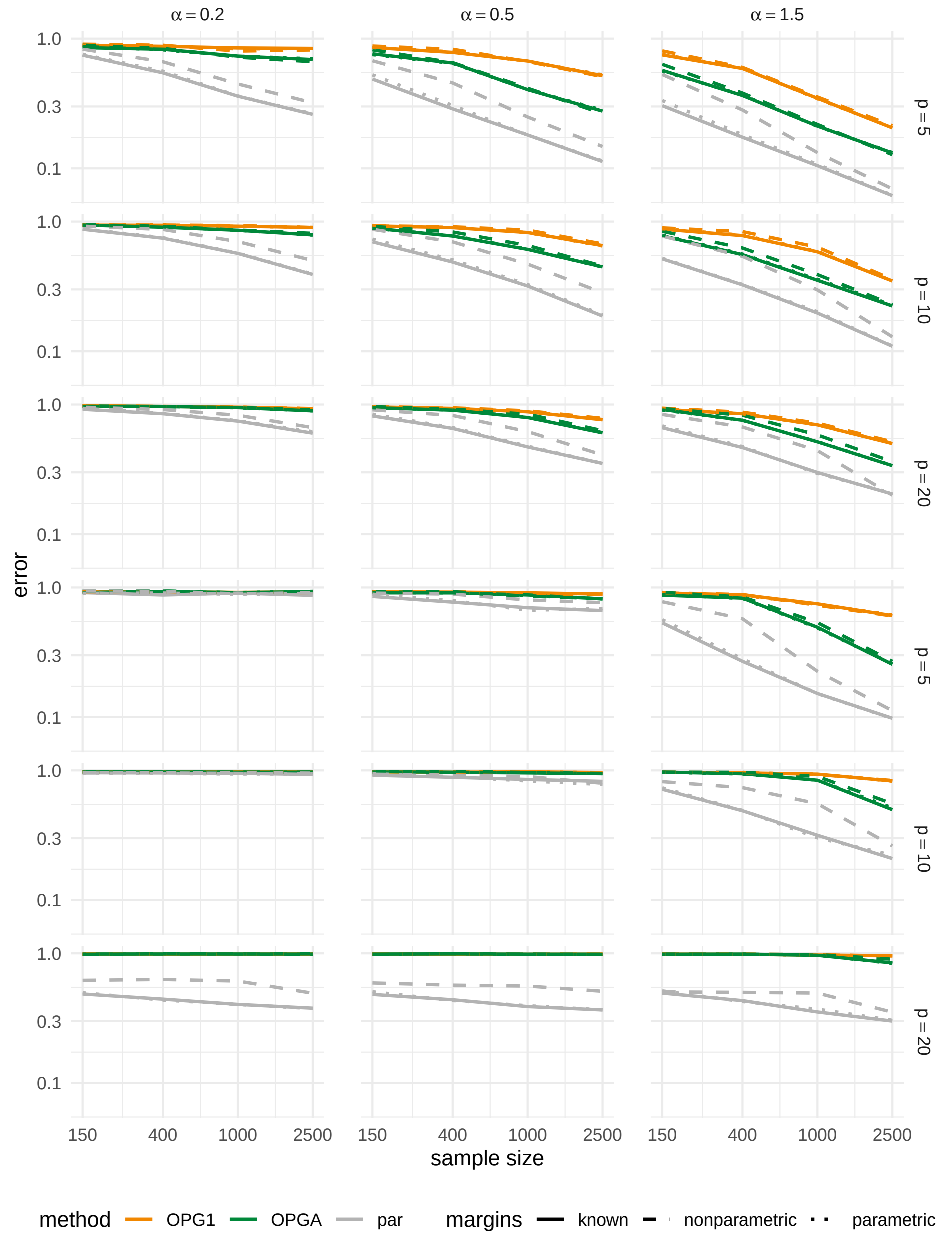

Fig. 2: Average error in estimating the central ρ subspace for different methods, sample sizes, signal strengths (α), covariate dimensions (p), and subspace dimensions (d). The conditional dependence is specified as a Gaussian copula. Both axes are on log-scale such that a linear trend with slope $-1/2$ corresponds to $\sqrt{n}$-convergence.

and $d = 2$ is an artifact of early termination of the likelihood optimization which leaves the final estimates closer to the initial values).

- The estimators under known and parametric margins perform essentially equally well. The nonparametric margins are slightly worse, but the difference is rather small.
- The error of the adaptive method appears to indeed achieve $\sqrt{n}$-convergence, at least in high-signal, large n settings. The sample size where the $\sqrt{n}$-convergence starts to kick in is larger for larger d and p.

Figure 3 provides additional results when varying the conditional dependence measure used for estimation and the family of the conditional copula in the case $\alpha = 0.5$, $p = 5$, and $d = 1$. Other configurations of dimension and signal strength are not shown, but lead to the same qualitative conclusions. We observe that the choice of copula family hardly affects the performance of OPG methods, but the parametric estimator is worse when the copula family is Clayton. Regarding the dependence measures, the performance is best for Spearman's ρ and Gini's γ. Estimation based on Blomqvist's β performs much worse. This is likely due to the fact that Blomqvist's β binarizes the data, leading to a loss in information. The van der Waerden's coefficient also performs slightly worse than the two best measures, potentially to the slightly heavier tails of the pseudo-response (sub-exponential instead of bounded).

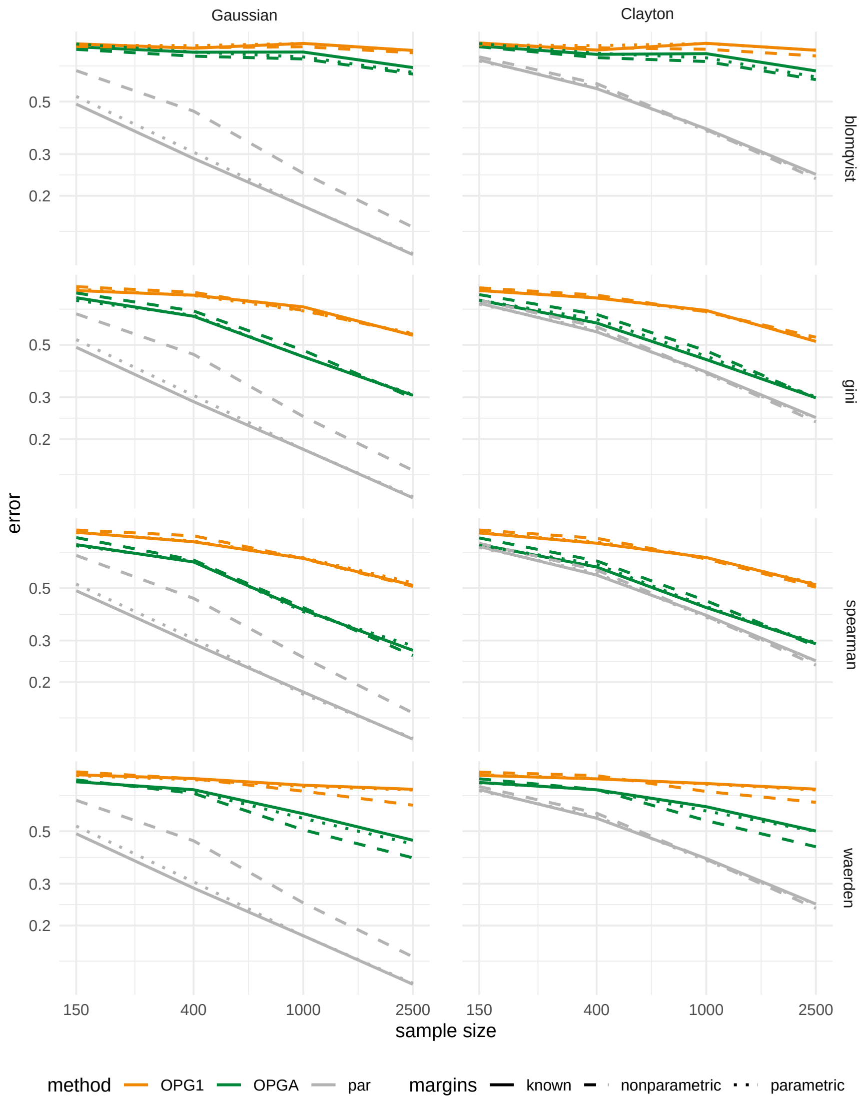

Fig. 3: Average error in estimating the central subspaces for different methods, sample sizes, dependence measures, and family of the conditional copula; here, $\alpha = 0.5$, $p = 5$, and $d = 1$. Both axes are on log-scale such that a linear trend with slope $-1/2$ corresponds to $\sqrt{n}$-convergence.

Appendix A Technical proofs

A.1 Preliminaries

Let $N(\epsilon, \mathcal{F}, L_1(P))$ be the minimal number of $L_1(P)$-balls of size ϵ required to cover a class of functions $\mathcal{F}$. A function F is called envelope of $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} |f(\boldsymbol{z})| \leq F(\boldsymbol{z})$ point-wise. For an arbitrary probability measure Q and function F, denote $\|f\|_Q = \int |f(\boldsymbol{z})| dQ(\boldsymbol{z})$.

Definition A.1 A class of functions $\mathcal{F}$ is called Euclidean with respect to an envelope F if there are constants $A, V < \infty$ such that $\sup_Q N(\epsilon\|F\|_Q, \mathcal{F}, L_1(Q)) \leq A\epsilon^{-V}$, where the supremum is over all probability measures with $\|F\|_Q > 0$.

Any class with a finite number of elements is trivially Euclidean with $V = 0$, but the classes can be much richer. A nice property of Euclidean classes is that new classes generates from sums and products of functions from Euclidean classes are also Euclidean (Pakes and Pollard 1989, Lemma 2.14). The following technical lemma will be useful later on.

Lemma A.1 *Let $r_1, \dots, r_p \in \mathbb{N}$, $\mathcal{D}$ be a bounded subset of $\mathbb{R}^p$, and K satisfy Assumption A1. For $\boldsymbol{s} \in \mathcal{D}$, define*

$$g_{\boldsymbol{x},\boldsymbol{B},h}(\boldsymbol{s}) = \prod_{i=1}^{p} (s_i - x_i)^{r_i} \times \prod_{i=1}^{p} K(\boldsymbol{B}_i^\top(\boldsymbol{s} - \boldsymbol{x})/h),$$

where $\boldsymbol{B}_i$ denotes the i-th column of $\boldsymbol{B} \in \mathbb{R}^{p\times p}$. Then the class of functions $\mathcal{F} = \left\{g_{\boldsymbol{x},\boldsymbol{B},h} \colon \boldsymbol{x} \in \mathcal{D}, \boldsymbol{B} \in \mathbb{R}^{p\times p}, h > 0\right\}$ is Euclidean with bounded envelope.

Proof The class of functions $\{s_k \mapsto s_k - x_k \colon x_k \in \mathcal{X}_k\}$ is Euclidean by Lemmas 2.4 and 2.12 of Pakes and Pollard (1989). Furthermore, the class

$$\{\boldsymbol{s} \mapsto K(\boldsymbol{B}_i(\boldsymbol{s} - \boldsymbol{x})/h) \colon \boldsymbol{B}_i \in \mathbb{R}^p, \boldsymbol{x} \in \mathbb{R}^p, h > 0\}$$

is Euclidean by Lemma 22 of Nolan and Pollard (1987). The elements of the class $\mathcal{F}$ are products of such functions and, hence, the class is also Euclidean by Lemma 2.14 of Pakes and Pollard (1989). Since all terms in the product are uniformly bounded, we have a bounded envelope. □

Euclidean classes are small enough for uniform laws of large numbers to hold. We shall repeatedly use the following result, which also gives a rate of convergence. It is a simplifed version of Theorem II.37 of Pollard (2012).

Proposition A.1 *For each $n \in \mathbb{N}$, let $\mathcal{F}_n$ be a Euclidean class of functions with bounded envelope F_n and constants not depending on n. It holds that*

$$\sup_{f\in\mathcal{F}_n} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{Z}_i) - \mathrm{E}[f(\boldsymbol{Z})] \right| = O_p\left(\sqrt{\frac{\sigma_n^2 \ln n}{n}}\right),$$

where $\sigma_n^2 = \max\left\{\ln n/n, \sup_{f\in\mathcal{F}_n} \mathrm{E}\{f^2(\boldsymbol{Z})\}\right\}$.

In everything that follows, we let $\boldsymbol{B} \in \mathbb{R}^{p\times p}, \boldsymbol{B}_0 \in \mathbb{R}^{p\times d}$ and define $\delta_{\boldsymbol{B}} = \|\boldsymbol{B} - (\boldsymbol{B}_0, \boldsymbol{0}_{p\times(p-d)})\|$, where $\|\cdot\|$ denotes the operator norm. The next two lemmas are concerned with kernel smoothing with approximately low-rank bandwidth matrices. For a d-dimensional vector $\boldsymbol{y} = (y_1, \dots, y_d)$ denote $\|\boldsymbol{y}\|_\infty = \max(|y_1, \dots, |y_d|)$.

Lemma A.2 *Suppose K satisfies A1 and $\mathcal{D}$ is a set with $A = 2\sup_{\boldsymbol{x}\in\mathcal{D}} \|\boldsymbol{x}\|_2 < \infty$. It holds that*

$$|K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) - K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)K(0)^{p-d}| = \mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \le h + A\delta_{\boldsymbol{B}}} \times O(\delta_{\boldsymbol{B}}/h),$$

uniformly in $\boldsymbol{x} \in \mathcal{D}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times p}$.

Proof Observe that $K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) = 0$ whenever $\|\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty > h$. Since for $A = 2\sup_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{X}}} \|\boldsymbol{x}\|_2$, we have $\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \le \|\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty + A\delta_{\boldsymbol{B}}$, it holds $K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) = K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) = 0$ whenever $\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty > h + A\delta_{\boldsymbol{B}}$. Furthermore, the Lipschitz property of the kernel implies

$$|K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) - K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)K(0)^{p-d}| = O(\delta_{\boldsymbol{B}}/h). \qquad \square$$

Lemma A.3 *Suppose A1–A2 hold, and g is a bounded function such that*

$$\alpha(\boldsymbol{z}) = \mathrm{E}\{g(\boldsymbol{X}) \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{z}\}$$

is twice continuously differentiable on $\{\boldsymbol{z} = \boldsymbol{B}_0^\top \boldsymbol{x} \colon \boldsymbol{x} \in \mathcal{D}\}$ for some compact set $\mathcal{D}$. If $h \to 0$ and $\delta_{\boldsymbol{B}}/h \to 0$, it holds uniformly in $\boldsymbol{x} \in \mathcal{D}$:

$$\begin{aligned} E[h^{p-d} g(\boldsymbol{X}) K_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] &= K(0)^{p-d} \alpha(\boldsymbol{B}_0^\top \boldsymbol{x}) f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}) + O(h^2 + \delta_{\boldsymbol{B}}/h), \\ \mathrm{Var}[h^{p-d} g(\boldsymbol{X}) K_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] &= O(h^{-d}). \end{aligned}$$

Proof By Lemma A.2, we have

$$K(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x})/h) = K(\boldsymbol{B}_0^\top(\boldsymbol{X} - \boldsymbol{x})/h)K(0)^{p-d} + \mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{\boldsymbol{x}}\|_\infty \le h + A\delta_{\boldsymbol{B}}} \times O(\delta_{\boldsymbol{B}}/h),$$

and, thus,

$$\begin{aligned} &\mathrm{E}[h^{p-d} g(\boldsymbol{X}) K_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] \\ &= h^{-d} \mathrm{E}[g(\boldsymbol{X}) K(\boldsymbol{B}_0^\top(\boldsymbol{X} - \boldsymbol{x})/h)] + h^{-d} \mathrm{E}[g(\boldsymbol{X}) \mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{\boldsymbol{x}}\|_\infty \le h + A\delta_{\boldsymbol{B}}}] \times O(\delta_{\boldsymbol{B}}/h) \\ &:= E_1 + E_2. \end{aligned}$$

The law of iterated expectations (first equality) and a change of variables (third equality) yield

$$\begin{aligned} E_1 &= h^{-d} K(0)^{p-d} \mathrm{E}[\alpha(\boldsymbol{B}_0^\top \boldsymbol{X}) K(\boldsymbol{B}_0^\top(\boldsymbol{X} - \boldsymbol{x})/h)] \\ &= h^{-d} K(0)^{p-d} \int \alpha(\boldsymbol{s}) K((\boldsymbol{s} - \boldsymbol{B}_0^\top \boldsymbol{x})/h) f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{s}) d\boldsymbol{s} \\ &= K(0)^{p-d} \int_{[-1,1]^d} \alpha(\boldsymbol{B}_0^\top \boldsymbol{x} - h\boldsymbol{t}) K(\boldsymbol{t}) f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x} - h\boldsymbol{t}) d\boldsymbol{t}. \end{aligned}$$

Expanding α and $f_{\boldsymbol{B}_0^\top \boldsymbol{X}}$ around $\boldsymbol{B}_0^\top \boldsymbol{x}$ and noting $\int K(s)ds = 1$ and $\int sK(s)ds = 0$, we obtain

$$E_1 = K(0)^{p-d}\alpha(\boldsymbol{B}_0^\top \boldsymbol{x}) f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}) + O(h^2).$$

Furthermore, since g is bounded and $\delta_{\boldsymbol{B}} = o(h)$, we have

$$\begin{aligned}
|E_2| &\leq h^{-d} |\mathrm{E}[|g(\boldsymbol{X})| \mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{\boldsymbol{x}}\|_\infty \leq h + A\delta_{\boldsymbol{B}}}] \times O(\delta_{\boldsymbol{B}}/h) \\
&\leq h^{-d} \times O(1) \times \mathrm{E}[\mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{\boldsymbol{x}}\|_\infty \leq h + A\delta_{\boldsymbol{B}}}] \times O(\delta_{\boldsymbol{B}}/h) \\
&= h^{-d} \times O(1) \times O((h + A\delta_{\boldsymbol{B}})^d) \times O(\delta_{\boldsymbol{B}}/h) \\
&= O(\delta_{\boldsymbol{B}}/h).
\end{aligned}$$

For the variance bound, we apply our result for expectation. Specifically, define $\widetilde{g}(\boldsymbol{x}) = g(\boldsymbol{x})^2$ and $\widetilde{K}(u) = K(u)^2/\kappa_2$, where $\kappa_2 = \int K(u)^2 du$. Observe that $\widetilde{g}$ satisfies the smoothness requirements of the lemma, and observe that $\widetilde{K}$ is also a bounded, Lipschitz continuous, symmetric probability density with support $[-1, 1]$. Noting $\widetilde{K}_h(\boldsymbol{x}) = h^{-p}\widetilde{K}(\boldsymbol{x}/h)$, we have

$$K_h(\boldsymbol{x})^2 = h^{-2p} K(\boldsymbol{x}/h)^2 = h^{-2p}\kappa_2^p \widetilde{K}(\boldsymbol{x}/h) = h^{-p}\kappa_2^p \widetilde{K}_h(\boldsymbol{x}).$$

Then

$$\begin{aligned}
&\mathrm{Var}[h^{p-d} g(\boldsymbol{X}) K_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] \leq \mathrm{E}[h^{2(p-d)} g(\boldsymbol{X})^2 K_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))^2] \\
&= \kappa_2^p \mathrm{E}[h^{p-2d} \widetilde{g}(\boldsymbol{X}) \widetilde{K}_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] = h^{-d}\kappa_2^p \mathrm{E}[h^{p-d} \widetilde{g}(\boldsymbol{X}) \widetilde{K}_h(\boldsymbol{B}^\top(\boldsymbol{X} - \boldsymbol{x}))] \\
&= h^{-d} \times O(1) = O(h^{-d}),
\end{aligned}$$

by the first part of the lemma. □

A.2 Uniform approximation of the local linear estimator

Let $\boldsymbol{A}^+$ denote the Moore-Penrose inverse of a matrix $\boldsymbol{A}$ and define $r_{n,d} = (nh^d/\ln n)^{-1/2}$. In this section, we prove the following result.

Theorem A.1 *Suppose that $d < p$, $\mathcal{G}$ and $\mathcal{H} = \{\boldsymbol{z} \mapsto \mathrm{E}\{g(\boldsymbol{Y}) \mid \boldsymbol{Z} = \boldsymbol{z}\} \colon g \in \mathcal{G}\}$ are Euclidean classes and A1–A4 hold. Then uniformly in $g \in \mathcal{G}, \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$, and all $\boldsymbol{B} \in \mathbb{R}^{p\times p}$ with $\delta_{\boldsymbol{B}}/h \to 0$, it holds*

$$\left| \begin{pmatrix} \widehat{m}_g(\boldsymbol{x}) \\ \widehat{\nabla} m_g(\boldsymbol{x}) \end{pmatrix} - \begin{pmatrix} m_g(\boldsymbol{x}) \\ \nabla m_g(\boldsymbol{x}) \end{pmatrix} - \boldsymbol{v}_{n,g}^*(\boldsymbol{x}) \right| = \begin{pmatrix} O(h^2) \\ O(h^4) \end{pmatrix} + O_p(h^2 r_{n,d} + h\delta_{\boldsymbol{B}} + r_{n,d}\delta_{\boldsymbol{B}}/h),$$

where

$$\begin{aligned}
&\boldsymbol{v}_{n,g}^*(\boldsymbol{x}) \\
&= \frac{1}{n f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x})} \sum_{i=1}^n \begin{pmatrix} \epsilon_g(\boldsymbol{Y}_i, \boldsymbol{X}_i)\big[1 + \{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}\}^\top \boldsymbol{\Gamma}^+(\boldsymbol{x})\{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{X}_i\}\big] K_h(\boldsymbol{B}_0{}^\top \boldsymbol{X}_{i,\boldsymbol{x}}) \\ \boldsymbol{\Gamma}^+(\boldsymbol{x})\epsilon_g(\boldsymbol{Y}_i, \boldsymbol{X}_i)\{\boldsymbol{X}_i - \boldsymbol{\mu}(\boldsymbol{x})\} K_h(\boldsymbol{B}_0{}^\top \boldsymbol{X}_{i,\boldsymbol{x}}) \end{pmatrix},
\end{aligned}$$

$$\begin{aligned}
\epsilon_g(\boldsymbol{Y}_i, \boldsymbol{X}_i) &= g(\boldsymbol{Y}_i) - \mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\}, \\
\boldsymbol{\mu}(\boldsymbol{x}) &= \mathrm{E}\{\boldsymbol{X} \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}\}, \\
\boldsymbol{\Gamma}(\boldsymbol{x}) &= \mathrm{E}\{\boldsymbol{X}\boldsymbol{X}^\top \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}\} - \boldsymbol{\mu}(\boldsymbol{x})\boldsymbol{\mu}(\boldsymbol{x})^\top.
\end{aligned}$$

The result follows from a sequence of lemmas stated and proven in the following. Equating the derivative of the local-linear estimation criterion in (7.1) to zero yields

$$\begin{pmatrix} \widehat{m}_g(\boldsymbol{x}) - m_g(\boldsymbol{x}) \\ \widehat{\nabla} m_g(\boldsymbol{x}) - \nabla m_g(\boldsymbol{x}) \end{pmatrix} = \boldsymbol{S}_{n,g}^{+}(\boldsymbol{x}) \boldsymbol{\tau}_{n,g}(\boldsymbol{x}), \tag{A1}$$

where

$$\begin{aligned} \boldsymbol{S}_{n,g}(\boldsymbol{x}) &= \frac{h^{p-d}}{n} \sum_{i=1}^{n} \begin{pmatrix} 1 & \boldsymbol{X}_{i,\boldsymbol{x}}^\top \\ \boldsymbol{X}_{i,\boldsymbol{x}} & \boldsymbol{X}_{i,\boldsymbol{x}} \boldsymbol{X}_{i,\boldsymbol{x}}^\top \end{pmatrix} K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}), \\ \boldsymbol{\tau}_{n,g}(\boldsymbol{x}) &= \frac{h^{p-d}}{n} \sum_{i=1}^{n} \left\{ g(\boldsymbol{Y}_i) - \boldsymbol{\beta}(\boldsymbol{x})^\top \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} \right\} \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}). \end{aligned} \tag{A2}$$

Lemma A.4 *If $h \to 0$, $\delta_{\boldsymbol{B}}/h \to 0$, and A1–A2 hold, then*

$$\sup_{\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}, \boldsymbol{B} \in \mathbb{R}^{p \times p}} \|\boldsymbol{S}_{n,g}(\boldsymbol{x}) - \boldsymbol{S}_g(\boldsymbol{x})\| = O_p(h^2 + \delta_{\boldsymbol{B}}/h + r_{n,d}),$$

where

$$\boldsymbol{S}_g(\boldsymbol{x}) = K(0)^{p-d} f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}) \mathrm{E}\left\{ \begin{pmatrix} 1 & \boldsymbol{X}_{\boldsymbol{x}}^\top \\ \boldsymbol{X}_{\boldsymbol{x}} & \boldsymbol{X}_{\boldsymbol{x}} \boldsymbol{X}_{\boldsymbol{x}}^\top \end{pmatrix} \middle| \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x} \right\}.$$

Proof Lemma A.3 implies that $\mathrm{E}\{\boldsymbol{S}_{n,g}(\boldsymbol{x})\} = \boldsymbol{S}_g(\boldsymbol{x}) + O(h^2 + \delta_{\boldsymbol{B}}/h)$, uniformly. By Lemma A.1, the classes of functions

$$\mathcal{R}_{j,k} = \left\{ \boldsymbol{s} \mapsto h^p \begin{pmatrix} 1 & (\boldsymbol{s} - \boldsymbol{x})^\top \\ (\boldsymbol{s} - \boldsymbol{x}) & (\boldsymbol{s} - \boldsymbol{x})(\boldsymbol{s} - \boldsymbol{x})^\top \end{pmatrix}_{j,k} K_h(\boldsymbol{B}^\top (\boldsymbol{s} - \boldsymbol{x})) \colon \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}, \boldsymbol{B} \in \mathbb{R}^{p \times p} \right\}$$

are Euclidean for any $j, k \in \{1, \dots, p+1\}$. By Lemma A.3, we also have

$$\max_{1 \le j,k \le p+1} \sup_{r \in \mathcal{R}_{j,k}} \mathrm{E}\{r(\boldsymbol{X})^2\} = O(h^d),$$

so that Proposition A.1 yields

$$\sup_{\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}, \boldsymbol{B} \in \mathbb{R}^{p \times p}} \|\boldsymbol{S}_{n,g}(\boldsymbol{x}) - \mathrm{E}\{\boldsymbol{S}_{n,g}(\boldsymbol{x})\}\| = O_p(r_{n,d}),$$

completing the proof. □

Corollary A.1 *Under A1–A3, $\boldsymbol{S}_{n,g}^{+}(\boldsymbol{x}) = \boldsymbol{S}_g^{+}(\boldsymbol{x})\{1 + o_p(1)\}$, uniformly on $\mathcal{D}_{\boldsymbol{X}}$.*

Lemma A.5 *It holds*

$$\boldsymbol{S}_g^{+}(\boldsymbol{x}) = \frac{1}{f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x})} \begin{pmatrix} 1 + \{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}\}^\top \boldsymbol{\Gamma}^{+}(\boldsymbol{x}) \{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}\} & -\{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}\}^\top \boldsymbol{\Gamma}^{+}(\boldsymbol{x}) \\ -\boldsymbol{\Gamma}^{+}(\boldsymbol{x}) \{\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}\} & \boldsymbol{\Gamma}^{+}(\boldsymbol{x}) \end{pmatrix},$$

Proof Recall block inversion formula

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{pmatrix},$$

which we can simplify to

$$\begin{pmatrix} 1 & \boldsymbol{C}^\top \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} 1 + \boldsymbol{C}^\top(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{C}^\top)^{-1}\boldsymbol{C} & -\boldsymbol{C}^\top(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{C}^\top)^{-1} \\ -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{C}^\top)^{-1}\boldsymbol{C} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{C}^\top)^{-1} \end{pmatrix}.$$

Then the result follows from setting

$$\boldsymbol{C} = \mathrm{E}(\boldsymbol{X}_{\boldsymbol{x}} \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x}, \quad \boldsymbol{D} = \mathrm{E}(\boldsymbol{X}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}}^\top \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}),$$

which gives

$$\begin{aligned}
\boldsymbol{D} - \boldsymbol{C}\boldsymbol{C}^\top &= \mathrm{E}(\boldsymbol{X}_{\boldsymbol{x}}\boldsymbol{X}_{\boldsymbol{x}}^\top \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}) \\
&\quad - \mathrm{E}(\boldsymbol{X}_{\boldsymbol{x}} \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x})\mathrm{E}(\boldsymbol{X}_{\boldsymbol{x}} \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x})^\top \\
&= \mathrm{E}[\boldsymbol{X}\boldsymbol{X}^\top \mid B_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}] - \boldsymbol{\mu}(\boldsymbol{x})\boldsymbol{x}^\top - \boldsymbol{x}\boldsymbol{\mu}(\boldsymbol{x})^\top - \boldsymbol{x}\boldsymbol{x}^\top \\
&\quad - (\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x})(\boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x})^\top. \\
&= \mathrm{E}[\boldsymbol{X}\boldsymbol{X}^\top \mid B_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}] - \boldsymbol{\mu}(\boldsymbol{x})\mu(\boldsymbol{x})^\top = \boldsymbol{\Gamma}(\boldsymbol{x}).
\end{aligned}$$

□

The structure of $\boldsymbol{S}_g^+(\boldsymbol{x})$ will cancel the leading bias term for the gradient estimate. To see this, we first derive an approximation of $\boldsymbol{\tau}_{n,g}(\boldsymbol{x})$ defined in (A2). For an arbitrary function s, denote $\partial_{i_1,\cdots,i_m} s(\boldsymbol{u}) = \partial^m s(\boldsymbol{u})/(\partial u_{i_1} \cdots \partial u_{i_m})$.

Lemma A.6 *Under the conditions of Theorem A.1, it holds*

$$\sup_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{X}}, g\in\mathcal{G}} \left\| \boldsymbol{\tau}_{n,g}(\boldsymbol{x}) - h^2\boldsymbol{\rho}(\boldsymbol{x}) - \boldsymbol{v}_{n,g}(\boldsymbol{x}) \right\| = O_p(h^4 + h^2 r_{n,d} + h\delta_{\boldsymbol{B}} + r_{n,d}\delta_{\boldsymbol{B}}/h),$$

uniformly in $g \in \mathcal{G}, \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$, and $\boldsymbol{B} \in \mathbb{R}^{p\times p}$ with $\delta_{\boldsymbol{B}}/h \to 0$, where

$$\begin{aligned}
\boldsymbol{\rho}(\boldsymbol{x}) &= \frac{K(0)^{p-d}}{2} \begin{pmatrix} 1 \\ \boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x} \end{pmatrix} f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}) \int_{[-1,1]^d} \boldsymbol{t}^\top \nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{x})\boldsymbol{t}K(\boldsymbol{t})d\boldsymbol{t} \\
\boldsymbol{v}_{n,g}(\boldsymbol{x}) &= \frac{1}{n}\sum_{i=1}^n \left[g(\boldsymbol{Y}_i) - \mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\}\right] \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} K_h({\boldsymbol{B}_0}^\top \boldsymbol{X}_{i,\boldsymbol{x}})K(0)^{p-d}.
\end{aligned}$$

Proof Decompose

$$\begin{aligned}
\boldsymbol{\tau}_{n,g}(\boldsymbol{x}) &= \frac{h^{p-d}}{n}\sum_{i=1}^n \left\{ g(\boldsymbol{Y}_i) - \boldsymbol{\beta}(\boldsymbol{x})^\top \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} \right\} \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}) \\
&= \frac{h^{p-d}}{n}\sum_{i=1}^n \left[g(\boldsymbol{Y}_i) - \mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\}\right] \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}) \\
&\quad + \frac{h^{p-d}}{n}\sum_{i=1}^n \left[\mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\} - \boldsymbol{\beta}(\boldsymbol{x})^\top \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} \right] \begin{pmatrix} 1 \\ \boldsymbol{X}_{i,\boldsymbol{x}} \end{pmatrix} K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}) \\
&= \boldsymbol{v}_{n,g}(\boldsymbol{x}, \boldsymbol{B}) + \boldsymbol{a}_{n,g}(\boldsymbol{x}, \boldsymbol{B}).
\end{aligned}$$

We deal with the two terms separately.

Term $\boldsymbol{v}_{n,g}(\boldsymbol{x}, \boldsymbol{B})$

It holds

$$
\begin{aligned}
&\boldsymbol{v}_{n,g}(\boldsymbol{x}, \boldsymbol{B}) \\
&= \boldsymbol{v}_{n,g}(\boldsymbol{x}) \\
&\quad + \frac{h^{-d}}{n}\sum_{i=1}^{n}\left[g(\boldsymbol{Y}_i) - \mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\}\right]\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}\left\{K\left(\frac{\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}}{h}\right) - K\left(\frac{\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}}{h}\right)K(0)^{p-d}\right\} \\
&= \boldsymbol{v}_{n,g}(\boldsymbol{x}) + \overline{\boldsymbol{v}}_{n,g}(\boldsymbol{x}, \boldsymbol{B}).
\end{aligned}
$$

The classes $\mathcal{G}$ and $\mathcal{H}$ are Euclidean by assumption. By Lemma A.1, also the classes

$$
\mathcal{V}_j = \left\{\boldsymbol{s} \mapsto h^p \begin{pmatrix}1\\ \boldsymbol{s}-\boldsymbol{x}\end{pmatrix}_j \{K_h(\boldsymbol{B}^\top(\boldsymbol{s}-\boldsymbol{x})) - K_h(\boldsymbol{B}_0{}^\top(\boldsymbol{s}-\boldsymbol{x}))\} \colon \boldsymbol{B} \in \mathbb{R}^{p\times p} \colon \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}\right\}
$$

are Euclidean. Since products of Euclidean classes are Euclidean, each coordinate of $h^d\overline{\boldsymbol{v}}_{n,g}(\boldsymbol{x}, \boldsymbol{B})$ is a sample average indexed by a Euclidean class. By iterated expectations, each element in the class has zero expectation. By Lemma A.2, each element is bounded by $A\mathbb{1}_{\|\boldsymbol{B}_0^\top(\boldsymbol{s}-\boldsymbol{x})\| \le h + A\delta_{\boldsymbol{B}}} \times O(\delta_{\boldsymbol{B}}/h)$, where $A = 2\sup_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{X}}}\|\boldsymbol{x}\|_2$. Since $f_{\boldsymbol{B}_0^\top \boldsymbol{X}}$ is bounded, we have $\mathrm{E}\left[\mathbb{1}^2_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\| \le h + A\delta_{\boldsymbol{B}}}\right] = O(h^d)$, it follows that

$$
\max_{1\le j\le p}\sup_{v\in\mathcal{V}_j}\mathrm{E}[v(\boldsymbol{X})^2/\delta_{\boldsymbol{B}}^2] = O_p(h^{d-2}).
$$

Again using Proposition A.1, we obtain

$$
\sup_{\boldsymbol{x}\in\mathcal{D}_{\boldsymbol{X}}, \boldsymbol{B}\in\mathbb{R}^{p\times p}, g\in\mathcal{G}}\|\overline{\boldsymbol{v}}_{n,g}(\boldsymbol{x}, \boldsymbol{B})\|/\delta_{\boldsymbol{B}} = O_p(r_{n,d}/h).
$$

Term $\boldsymbol{a}_{n,g}(\boldsymbol{x}, \boldsymbol{B})$

Define

$$
\gamma(\boldsymbol{X}_i, \boldsymbol{x}) = \mathrm{E}\{g(\boldsymbol{Y}_i) \mid \boldsymbol{X}_i\} - \boldsymbol{\beta}(\boldsymbol{x})^\top \begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix},
$$

and observe that

$$
\gamma(\boldsymbol{X}_i, \boldsymbol{x}) = m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{X}_i) - m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{x}) - \nabla m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{x})\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}.
$$

In fact, $\gamma(\boldsymbol{X}_i, \boldsymbol{x})$ can be written as a function of $\boldsymbol{B}_0^\top \boldsymbol{X}_i$ and $\boldsymbol{B}_0^\top \boldsymbol{x}$ only, and we write $\gamma(\boldsymbol{X}_i, \boldsymbol{x}) = \gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)$ in what follows. With this notation, we have

$$
\boldsymbol{a}_{n,g}(\boldsymbol{x}, \boldsymbol{B}) = h^2\frac{h^{p-d}}{n}\sum_{i=1}^{n}\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}).
$$

We compute the expectation and variance of the summands similarly to the proof of Lemma A.3. We have

$$
\begin{aligned}
&\mathrm{E}\left[h^{p-d}\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}})\right] \\
&= h^{-d}\mathrm{E}\left[\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)K(0)^{p-d}\right] \\
&\qquad + h^{-d}\mathrm{E}\left[\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}\{K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) - K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)K(0)^{p-d}\}\right]
\end{aligned}
$$

$$= E_1 + E_2.$$

For E_1, the law of iterated expectations implies

$$E_1 = K(0)^{p-d}\mathrm{E}\left[\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{\mu}(\boldsymbol{X}_i)-\boldsymbol{x}\end{pmatrix}K_h(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}})\right].$$

Define $\widetilde{\boldsymbol{\mu}}(\boldsymbol{B}_0^\top \boldsymbol{x}) = \mathrm{E}[\boldsymbol{X} \mid \boldsymbol{B}_0^\top \boldsymbol{X} = \boldsymbol{B}_0^\top \boldsymbol{x}] = \boldsymbol{\mu}(\boldsymbol{x})$. The change of variables $\boldsymbol{t} = (\boldsymbol{s} - \boldsymbol{B}_0^\top \boldsymbol{x})/h$ gives

$$\begin{aligned}
&E_1/K(0)^{p-d}\\
&= \int h^{-2}\gamma_{\boldsymbol{x}}(\boldsymbol{s})\begin{pmatrix}1\\ \widetilde{\boldsymbol{\mu}}(\boldsymbol{s})-\boldsymbol{x}\end{pmatrix}K_h(\boldsymbol{s}-\boldsymbol{B}_0^\top \boldsymbol{x})f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{s})d\boldsymbol{s}\\
&= \int_{[-1,1]^d} h^{-2}\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{x}-h\boldsymbol{t})\begin{pmatrix}1\\ \widetilde{\boldsymbol{\mu}}(\boldsymbol{B}_0^\top \boldsymbol{x}-h\boldsymbol{t})-\boldsymbol{x}\end{pmatrix}K(\boldsymbol{t})f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x}-h\boldsymbol{t})d\boldsymbol{t}.
\end{aligned}$$

Observe that

$$\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{x}) = 0, \quad \nabla\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{x}) = \boldsymbol{0}, \quad \nabla^2\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{z}) = \nabla^2 m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{z}),$$

so that a fourth order Taylor expansion of $\gamma_{\boldsymbol{x}}$ around $\boldsymbol{B}_0^\top \boldsymbol{x}$ yields

$$h^{-2}\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{x}-h\boldsymbol{t}) = \frac{1}{2}\boldsymbol{t}^\top\nabla^2 m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{z})\boldsymbol{t} + \frac{h}{6}\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d t_it_jt_k\partial_{i,j,k}m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{x}) + O(h^2\|\boldsymbol{t}\|^4),$$

where $\partial_{i,j,k}m_g^{\boldsymbol{B}_0}(\boldsymbol{s}) = \partial^3 m_g^{\boldsymbol{B}_0}(\boldsymbol{s})/(\partial s_i\partial s_j\partial s_k)$. Also expanding $\widetilde{\boldsymbol{\mu}}$, and $f_{\boldsymbol{B}_0^\top \boldsymbol{X}}$ around $\boldsymbol{B}_0^\top \boldsymbol{x}$ (up to second order) and noting $\int K(s)ds = 1$ and $\int sK(s)ds = 0$, we obtain

$$\begin{aligned}
E_1 &= \frac{K(0)^{p-d}}{2}\begin{pmatrix}1\\ \widetilde{\boldsymbol{\mu}}(\boldsymbol{B}_0^\top \boldsymbol{x})-\boldsymbol{x}\end{pmatrix}f_{\boldsymbol{B}_0^\top \boldsymbol{X}}(\boldsymbol{B}_0^\top \boldsymbol{x})\int_{[-1,1]^d}\boldsymbol{t}^\top\nabla^2 m_g^{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top \boldsymbol{x})\boldsymbol{t}K(\boldsymbol{t})d\boldsymbol{t} + O(h^2)\\
&= \boldsymbol{\rho}(\boldsymbol{x}) + O(h^2).
\end{aligned}$$

Next, recall

$$E_2 = h^{-d}\mathrm{E}\left[\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}\{K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h) - K(\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)K(0)^{p-d}\}\right].$$

By Lemma A.2, we have

$$|E_2| \le h^{-d}(1+2A)\mathrm{E}\left[\left|\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\right|\mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \le h+A\delta_{\boldsymbol{B}}}\right] \times O(\delta_{\boldsymbol{B}}/h).$$

Further, $\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \le h + A\delta_{\boldsymbol{B}}$ and $\delta_{\boldsymbol{B}} = o(h)$ imply $|\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)|/h = O(1)$. Thus,

$$|E_2| \le h^{-d}(1+2A)\times O(1)\times \mathrm{E}\left[\mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \le h+A\delta_{\boldsymbol{B}}}\right] \times O(\delta_{\boldsymbol{B}}/h) = O(\delta_{\boldsymbol{B}}/h).$$

Putting everything together, we have shown

$$\mathrm{E}[\boldsymbol{a}_{n,g}(\boldsymbol{x},\boldsymbol{B})] = h^2(E_1+E_2) = O(h^2 + \delta_{\boldsymbol{B}}h).$$

For the variance, we have

$$\begin{aligned}
&\left\|\mathrm{Var}\left[h^{p-d}\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\begin{pmatrix}1\\ \boldsymbol{X}_{i,\boldsymbol{x}}\end{pmatrix}K_h(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}})\right]\right\|\\
&\le h^{-2d}2(1+A^2)\mathrm{E}\left[\left|\frac{\gamma_{\boldsymbol{x}}(\boldsymbol{B}_0^\top \boldsymbol{X}_i)}{h^2}\right|^2 K(\boldsymbol{B}^\top \boldsymbol{X}_{i,\boldsymbol{x}}/h)^2\right]
\end{aligned}$$

$$\leq O(h^{-2d}) \times \mathrm{E}\left[\mathbb{1}_{\|\boldsymbol{B}_0^\top \boldsymbol{X}_{i,\boldsymbol{x}}\|_\infty \leq h + A\delta_{\boldsymbol{B}}}\right] = O(h^{-d}),$$

by the same arguments as in Lemma A.3. Using Proposition A.1, we obtain

$$\|\boldsymbol{a}_{n,g}(\boldsymbol{x}, \boldsymbol{B}) - \mathrm{E}[\boldsymbol{a}_{n,g}(\boldsymbol{x}, \boldsymbol{B})]\| = O_p(r_{n,d}h^2 + \delta_{\boldsymbol{B}} h),$$

uniformly for $g \in \mathcal{G}, \boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$, $\boldsymbol{B} \in \mathbb{R}^{p \times p}$ with $\delta_{\boldsymbol{B}}/h \to 0$. This completes the proof. □

Now we see that all components of $\boldsymbol{S}_g^+(\boldsymbol{x})\boldsymbol{\rho}(\boldsymbol{x})$ except the first cancel out. This completes the proof of Theorem A.1.

A.3 Proof of Theorem 7.1

For $t = 0$, arguments similar to Lemma A.4 and Lemma A.6 with $p = d$ and $\boldsymbol{B}_0 = \widehat{\boldsymbol{B}}^{(0)}$ yield $\|\widehat{\boldsymbol{\Delta}}_{\mathcal{G}} - \boldsymbol{\Delta}_{\mathcal{G}}\| = O_p(h^2 + r_{n,p}/h)$. By Assumption A5 and Yu et al. (2015, Theorem 2),

$$\delta_{\widehat{\boldsymbol{B}}^{(0)}} = \|\widehat{\boldsymbol{B}}^{(0)} - \boldsymbol{B}_0\| \lesssim \|\widehat{\boldsymbol{\Delta}}_{\mathcal{G}} - \boldsymbol{\Delta}_{\mathcal{G}}\| = O_p(h_0^2 + r_{n,p}/h_0) = O_p\{(\ln n/n)^{2/(6+p)}\}.$$

Now consider $t > 0$ and define $r_{n,d,t} = (nh_t^d/\ln n)^{-1/2}$, $\sigma_{n,t} = h_t^4 + h_t^2 r_{n,d,t} + h_t \delta_{\widehat{\boldsymbol{B}}^{(t)}} + r_{n,d,t}\delta_{\widehat{\boldsymbol{B}}^{(t)}}/h_t$. First expand

$$\begin{aligned}
\widehat{\boldsymbol{\Delta}}_g &= \frac{1}{n}\sum_{i=1}^n \widehat{\nabla} m_g(\boldsymbol{X}_i)\widehat{\nabla} m_g(\boldsymbol{X}_i)^\top \\
&= \frac{1}{n}\sum_{i=1}^n \nabla m_g(\boldsymbol{X}_i)\nabla m_g(\boldsymbol{X}_i)^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^n \nabla m_g(\boldsymbol{X}_i)\{\widehat{\nabla} m_g(\boldsymbol{X}_i) - \nabla m_g(\boldsymbol{X}_i)\}^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{\widehat{\nabla} m_g(\boldsymbol{X}_i) - \nabla m_g(\boldsymbol{X}_i)\}\nabla m_g(\boldsymbol{X}_i)^\top \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{\widehat{\nabla} m_g(\boldsymbol{X}_i) - \nabla m_g(\boldsymbol{X}_i)\}\{\widehat{\nabla} m_g(\boldsymbol{X}_i) - \nabla m_g(\boldsymbol{X}_i)\}^\top \\
&= D_1 + D_2 + D_3 + D_4 = D_1 + D_2 + D_3 + O_p(\sigma_{n,t}^2 + r_{n,d,t}^2).
\end{aligned}$$

D_1 is an oracle version of $\widehat{\boldsymbol{\Delta}}$ and $D_1 = \boldsymbol{\Delta}_g + O_p(n^{-1/2})$. The other terms come from estimating ∇m_g.

Recall from Theorem A.1

$$\sup_{\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}, g \in \mathcal{G}} \left\| \widehat{\nabla} m_g(\boldsymbol{x}) - \nabla m_g(\boldsymbol{x}) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}_n(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{x}) \right\| = O_p(\sigma_{n,t}),$$

where

$$\boldsymbol{\psi}_n(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{x}) = \frac{1}{f_{\boldsymbol{B}_0^\top}(\boldsymbol{B}_0^\top \boldsymbol{x})}\boldsymbol{\Gamma}^+(\boldsymbol{x})\epsilon_g(\boldsymbol{Y}_i, \boldsymbol{X}_i)\big\{\boldsymbol{X}_i - \boldsymbol{\mu}(\boldsymbol{x})\big\}K_{h_t}(\boldsymbol{B}_0{}^\top \boldsymbol{X}_{i,\boldsymbol{x}}).$$

Then,

$$
\begin{aligned}
D_2 = D_3^\top &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \nabla m_g(\boldsymbol{X}_i)\boldsymbol{\psi}_n(\boldsymbol{Y}_j,\boldsymbol{X}_j,\boldsymbol{X}_i)^\top + O_p(\sigma_{n,t}) \\
&= \binom{n}{2}^{-2}\sum_{i=1}^n\sum_{j=i+1}^n \boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_j,\boldsymbol{X}_j) + O_p(\sigma_{n,t}+n^{-1}),
\end{aligned}
$$

where

$$
\boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_j,\boldsymbol{X}_j) = \nabla m_g(\boldsymbol{X}_i)\boldsymbol{\psi}_n(\boldsymbol{Y}_j,\boldsymbol{X}_j,\boldsymbol{X}_i)^\top + \nabla m_g(\boldsymbol{X}_j)\boldsymbol{\psi}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{X}_j)^\top.
$$

Since $\mathrm{E}\{\epsilon_g(\boldsymbol{Y}_j,\boldsymbol{X}_j)\mid\boldsymbol{X}_j\}=0$, it holds $\mathrm{E}\{\boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{X}_j)\}=0$ and

$$
\begin{aligned}
&\mathrm{E}\{\boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{X}_j)\mid\boldsymbol{Y}_i,\boldsymbol{X}_i\} \\
&= \mathrm{E}\{\nabla m_g(\boldsymbol{X}_j)\boldsymbol{\psi}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{X}_j)^\top\mid\boldsymbol{Y}_i,\boldsymbol{X}_i\} \\
&= \epsilon(\boldsymbol{Y}_i,\boldsymbol{X}_i)\mathrm{E}\big[\nabla m_g(\boldsymbol{X}_j)\boldsymbol{\Gamma}^+(\boldsymbol{X}_j)\{\boldsymbol{X}_i-\boldsymbol{\mu}(\boldsymbol{X}_j)\}K_{h_t}\{\boldsymbol{B}_0^\top(\boldsymbol{X}_i-\boldsymbol{X}_j)\}/f_{\boldsymbol{B}_0^\top}(\boldsymbol{B}_0^\top\boldsymbol{X}_j)\mid\boldsymbol{X}_i\big] \\
&= \epsilon(\boldsymbol{Y}_i,\boldsymbol{X}_i)\big[\nabla m_g(\boldsymbol{X}_i)\{\boldsymbol{X}_i-\boldsymbol{\mu}(\boldsymbol{X}_i)\}^\top\boldsymbol{\Gamma}^+(\boldsymbol{X}_i)+O(h_t^2)\big] \\
&=: \tilde{\boldsymbol{A}}(\boldsymbol{Y}_i,\boldsymbol{X}_i).
\end{aligned}
$$

Computing $\mathrm{Var}\big\{\boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_j,\boldsymbol{X}_j)-\tilde{\boldsymbol{A}}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i)-\tilde{\boldsymbol{A}}_n(\boldsymbol{Y}_j,\boldsymbol{X}_j)\big\}=O(h_t^{-d})$, standard results for U-statistics (e.g., Korolyuk and Borovskich 2013) yield

$$
\begin{aligned}
\binom{n}{2}^{-2}\sum_{i=1}^n\sum_{j=i+1}^n \boldsymbol{A}_n(\boldsymbol{Y}_i,\boldsymbol{X}_i,\boldsymbol{Y}_j,\boldsymbol{X}_j) &= \frac{1}{n}\sum_{i=1}^n \tilde{\boldsymbol{A}}(\boldsymbol{Y}_i,\boldsymbol{X}_i) + O_p(n^{-1}h_t^{-d/2}) \\
&= O_p(n^{-1/2}+r_{n,d}^2).
\end{aligned}
$$

We have shown that, $\|\widehat{\boldsymbol{\Delta}}_g-\boldsymbol{\Delta}_g\| = O_p(n^{-1/2}+\sigma_{n,t}+r_{n,d,t}^2)$. Then

$$
\begin{aligned}
\delta_{\widehat{\boldsymbol{B}}^{(t)}} &= O_p(n^{-1/2}+\sigma_{n,t}+r_{n,d,t}^2) \\
&= O_p(n^{-1/2}+h_t^3+h_t^2 r_{n,d,t}+h_t\delta_{\boldsymbol{B}^{(t-1)}}+r_{n,d,t}\delta_{\widehat{\boldsymbol{B}}^{(t)}}/h_t+r_{n,d,t}^2) \\
&= O_p\{n^{-1/2}+h_t^3+r_{n,d,t}^2+\delta_{\widehat{\boldsymbol{B}}^{(t-1)}}(h_0+r_{n,d,\infty}/h_\infty)\},
\end{aligned}
$$

which we write as $O_p(a_{n,t}+\delta_{\widehat{\boldsymbol{B}}^{(t-1)}}b_n)$ for simplicity. Iteratively substituting this expression for $\delta_{\widehat{\boldsymbol{B}}^{(t-1)}}$ yields

$$
\delta_{\widehat{\boldsymbol{B}}^{(t+1)}} = O_p\bigg\{a_{n,t}+\sum_{k=1}^{t-1}a_{n,t-k}b_n^k+\delta_{\widehat{\boldsymbol{B}}^{(0)}}b_n^t\bigg\}.
$$

Since $h_t = \max\{\rho h_{t-1}, h_\infty\}$ with $\rho \in (0,1)$, it holds $a_{n,t-k} = O(a_{n,t})$ for all $k \leq k^*$ and fixed $k^* < \infty$. Since also $b_n = o(1)$,

$$\delta_{\widehat{\boldsymbol{B}}^{(t+1)}} = O_p\left\{a_{n,t} + \sum_{k=k^*}^{t-1} a_{n,t-k} b_n^k + \delta_{\widehat{\boldsymbol{B}}^{(0)}} b_n^t\right\}.$$

Furthermore, there is a finite k^* such that $b_n^{k^*} = O(a_{n,t})$. Hence, for $t \geq k^*$, $\delta_{\widehat{\boldsymbol{B}}^{(t+1)}} = O_p(a_{n,t}) = O_p(n^{-1/2} + h_t^4 + r_{n,d,t}^2)$. Iterating until $h_t = h_\infty$ proves our claim. □

References

Acar, E.F., Craiu, R.V., Yao, F.: Dependence calibration in conditional copulas: A nonparametric approach. Biometrics **67**(2), 445–453 (2011)

Abegaz, F., Gijbels, I., Veraverbeke, N.: Semiparametric estimation of conditional copulas. Journal of Multivariate Analysis **110**, 43–73 (2012). Special Issue on Copula Modeling and Dependence

Chiaromonte, F., Cook, R.D.: On foundations of regression graphics. Technical Report 616, University of Minnesota (1997)

Cook, R.D., Li, B.: Dimension reduction for conditional mean in regression. The Annals of Statistics, 455–474 (2002)

Cook, R.D.: On the interpretation of regression plots. Journal of the American Statistical Association **89**(425), 177–189 (1994)

Cook, R.D.: Regression Graphics: Ideas for Studying Regressions Through Graphics. Probability and statistics section. Wiley, ??? (1998)

Cook, R.D., Setodji, C.M.: A model-free test for reduced rank in multivariate regression. Journal of the American Statistical Association **98**(462), 340–351 (2003)

Derumigny, A., Fermanian, J.-D.: About tests of the "simplifying" assumption for conditional copulas. Dependence Modeling **5**(1), 154–197 (2017)

Fermanian, J.-D., Lopez, O.: Single-index copulas. Journal of Multivariate Analysis **165**, 27–55 (2018)

Gijbels, I., Omelka, M., Veraverbeke, N., *et al.*: Partial and average copulas and association measures. Electronic Journal of Statistics **9**(2), 2420–2474 (2015)

Gijbels, I., Omelka, M., Veraverbeke, N.: Estimation of a copula when a covariate affects only marginal distributions. Scandinavian Journal of Statistics **42**(4), 1109–1126 (2015)

Gijbels, I., Omelka, M., Veraverbeke, N.: Nonparametric testing for no covariate effects in conditional copulas. Statistics **51**(3), 475–509 (2017)

Gijbels, I., Veraverbeke, N., Omelka, M.: Conditional copulas, association measures and their applications. Computational Statistics & Data Analysis **55**(5), 1919–1932 (2011)

Haffenden, H., Artemiou, A.: Using sliced inverse mean difference for dimension reduction in multivariate time series. Stat **13**(3), 709 (2024) https://doi.org/10.1002/sta4.709 https://onlinelibrary.wiley.com/doi/pdf/10.1002/sta4.709

Haff, I.H., Aas, K., Frigessi, A.: On the simplified pair-copula construction—simply useful or too simplistic? Journal of Multivariate Analysis **101**(5), 1296–1310 (2010)

Hausdorff, F.: Summationsmethoden und momentfolgen. i. Mathematische Zeitschrift **9**(1), 74–109 (1921)

Huang, M.-Y., Chiang, C.-T.: An effective semiparametric estimation approach for the sufficient dimension reduction model. Journal of the American Statistical Association **112**(519), 1296–1310 (2017) https://doi.org/10.1080/01621459.2016.1215987 https://doi.org/10.1080/01621459.2016.1215987

Härdle, W., Stoker, T.M.: Investigating smooth multiple regression by the method of average derivatives. Journal of the American Statistical Association **84**(408), 986–995 (1989)

Haff, I.H., Segers, J.: Nonparametric estimation of pair-copula constructions with the empirical pair-copula. Computational Statistics & Data Analysis **84**, 1–13 (2015)

Helmlinger, G., Yuan, F., Dellian, M., Jain, R.K.: Interstitial ph and po2 gradients in solid tumors in vivo: high-resolution measurements reveal a lack of correlation. Nature medicine **3**(2), 177–182 (1997)

Korolyuk, V.S., Borovskich, Y.V.: Theory of U-statistics vol. 273. Springer, ??? (2013)

Li, B., Artemiou, A., Li, L.: Principal support vector machines for linear and nonlinear sufficient dimension reduction. The Annals of Statistics **39**(6), 3182–3210 (2011)

Luo, W., Li, B., Yin, X., *et al.*: On efficient dimension reduction with respect to a statistical functional of interest. The Annals of Statistics **42**(1), 382–412 (2014)

Ma, Y., Zhu, L.: A semiparametric approach to dimension reduction. Journal of the American Statistical Association **107**(497), 168–179 (2012) https://doi.org/10.1080/01621459.2011.646925

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

Nagler, T.: Simplified vine copula models: state of science and affairs. arXiv preprint arXiv:2410.16806 (2024)

Nagler, T., Czado, C.: Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. Journal of Multivariate Analysis **151**, 69–89 (2016)

Nelsen, R.B.: An Introduction to Copulas. Springer Series in Statistics. Springer, ??? (2006)

Nolan, D., Pollard, D.: U-processes: rates of convergence. The Annals of Statistics, 780–799 (1987)

Patton, A.J.: Modelling time-varying exchange rate dependence using the conditional copula. UCSD Discussion Paper (2001)

Pollard, D.: Convergence of Stochastic Processes. Springer, ??? (2012)

Pakes, A., Pollard, D.: Simulation and the asymptotics of optimization estimators. Econometrica: Journal of the Econometric Society, 1027–1057 (1989)

Portier, F., Segers, J.: On the weak convergence of the empirical conditional copula under a simplifying assumption. Journal of Multivariate Analysis **166**, 160–181 (2018)

Rényi, A.: On measures of dependence. Acta mathematica hungarica **10**(3-4), 441–451 (1959)

Rodriguez, J.C.: Measuring financial contagion: A copula approach. Journal of Empirical Finance **14**(3), 401–423 (2007) https://doi.org/10.1016/j.jempfin.2006.07.002

Stoeber, J., Joe, H., Czado, C.: Simplified pair copula constructions—limitations and extensions. Journal of Multivariate Analysis **119**, 101–118 (2013)

Spanhel, F., Kurz, M.S., *et al.*: Simplified vine copula models: Approximations based on the simplifying assumption. Electronic Journal of Statistics **13**(1), 1254–1291 (2019)

Sklar, A.: Fonctions de Répartition À N Dimensions et Leurs marges. Université Paris 8, ??? (1959)

Schweizer, B., Wolff, E.F.: On nonparametric measures of dependence for random variables. The Annals of Statistics **9**(4), 879–885 (1981)

Vatter, T., Chavez-Demoulin, V.: Generalized additive models for conditional dependence structures. Journal of Multivariate Analysis **141**, 147–167 (2015)

Veraverbeke, N., Omelka, M., Gijbels, I.: Estimation of a conditional copula and association measures. Scandinavian Journal of Statistics **38**(4), 766–780 (2011)

Xia, Y.: A constructive approach to the estimation of dimension reduction directions. Ann. Statist. **35**(6), 2654–2690 (2007) https://doi.org/10.1214/009053607000000352

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X.: An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(3), 363–410 (2002) https://doi.org/10.1111/1467-9868.03411

Yin, X., Cook, R.D.: Dimension reduction for the conditional kth moment in regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64**(2), 159–175 (2002) https://doi.org/10.1111/1467-9868.00330

Yin, X., Li, B.: Sufficient dimension reduction based on an ensemble of minimum average variance estimators. The Annals of Statistics **39**(6), 3392–3416 (2011) https://doi.org/10.1214/11-AOS950

Yu, Y., Wang, T., Samworth, R.J.: A useful variant of the Davis–Kahan theorem for statisticians. Biometrika **102**(2), 315–323 (2015)

Zhang, L., Singh, V.: Bivariate flood frequency analysis using the copula method. Journal of Hydrologic Engineering **11**(2), 150–164 (2006)

Zhang, Q., Xue, L., Li, B.: Dimension reduction for fréchet regression. Journal of the American Statistical Association **119**(548), 2733–2747 (2024) https://doi.org/10.1080/01621459.2023.2277406

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