Journal of Machine Learning Research 23 (2022) 1-49 Submitted 4/21; Revised 11/21; Published 1/22

The correlation-assisted missing data estimator

Timothy I. Cannings TIMOTHY.CANNINGS@ED.AC.UK
School of Mathematics
University of Edinburgh
Edinburgh, UK.

Yingying Fan FANYINGY@MARSHALL.USC.EDU
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
Los Angeles, CA 90089, USA

Editor: Philipp Hennig

Abstract

We introduce a novel approach to estimation problems in settings with missing data. Our proposal – the *Correlation-Assisted Missing data* (CAM) estimator – works by exploiting the relationship between the observations with missing features and those without missing features in order to obtain improved prediction accuracy. In particular, our theoretical results elucidate general conditions under which the proposed CAM estimator has lower mean squared error than the widely used complete-case approach in a range of estimation problems. We showcase in detail how the CAM estimator can be applied to U-Statistics to obtain an unbiased, asymptotically Gaussian estimator that has lower variance than the complete-case U-Statistic. Further, in nonparametric density estimation and regression problems, we construct our CAM estimator using kernel functions, and show it has lower asymptotic mean squared error than the corresponding complete-case kernel estimator. We also include practical demonstrations throughout the paper using simulated data and the *Terneuzen birth cohort* and *Brandsma* datasets available from `CRAN`.

Keywords: Missing data, U-Statistics, kernel density estimation, local constant regression, nonparametric

1. Introduction

Data is a primary commodity in today's economy, it is valued and traded like any other asset. Statistics and machine learning allow us to extract this value by improving operational efficiency, increasing revenue, or understanding the behaviour of customers. A common complication in modern applications is that the data may be incomplete. For example, some users may choose not to disclose their personal details (age, gender, geographic location, etc.) to a smartphone application; optional questions on an on-line form are often left blank; or data is sometimes removed or hidden to guarantee privacy. In other situations, missing data problems can arise when two or more different data sources have been combined.

Missing data is not a new problem. As early as the 1950s, Anderson (1957) found the maximum likelihood estimator in a multivariate normal distribution when some of the observations were missing. In a seminal paper, Rubin (1976) studied missing data in a

rigorous general framework, introduced the notion of data *missing at random*, and specified conditions under which the process that causes data to be missing may be ignored. See also the comprehensive book on the subject by Little and Rubin (2002) and the recent special issue of *Statistical Science* (Josse and Reiter, 2018).

A simple and widely used approach to deal with missing data is to discard any incomplete observations – a technique referred to as complete-case analysis (Little and Rubin, 2002, Chapter 3). There is an obvious drawback with this method that perhaps much of the data is ignored. An alternative approach is to impute the missing values (Ford, 1983). There is an extensive body of work on different imputation techniques; see, for instance, Little and Rubin (2002, Chapters 4 and 5) and Molenberghs et al. (2015, Chapter IV) for an overview. Other techniques are based on the expectation–maximisation (EM) algorithm (Dempster et al., 1977). Another important line of work is known as *doubly-robust* estimation (Tsiatis, 2006), which combines the inverse probability weighted estimator with a bias correction. The term double robustness here refers to the fact that the method is consistent if either the missingness model or the regression model is correctly specified; see Kang and Schafer (2007) for an accessible overview. Missing data has also been studied in a range of high-dimensional settings, including regression (Loh and Wainwright, 2012), covariance matrix estimation (Lounici, 2014; Cai and Zhang, 2016), classification (Cai and Zhang, 2018), and (sparse) principal component analysis (Elsener and van de Geer, 2018; Zhu et al., 2019).

In this paper, we develop a novel approach to missing data problems. Our new proposal, the correlation-assisted missing data (CAM) estimator, exploits the relationship between the complete cases and the observations with missing values, in order to improve on the performance of the complete-case estimator. More precisely, we construct an (approximately) mean-zero statistic, using both the complete-cases and the data with missing entries, which is correlated with the complete-case estimator. We then exploit this correlation to construct our new estimator, by making a linear adjustment to the complete-case estimator.

Our first main result, Proposition 1, elucidates when the proposed CAM estimator will be more accurate than the complete-case estimator in terms of mean squared error. The result does not require any assumptions on the data generating mechanism. In particular, we do not assume the data to be missing completely at random. Further, Proposition 1 motivates an optimal (but typically unknown) choice of the adjustment term used in the construction of the CAM estimator. This optimal choice leads to the greatest reduction in mean squared error. We also show that we can then construct a data-driven version of the CAM estimator that performs well in many practical settings.

As a second main contribution, we showcase how the CAM estimation technique can be applied in specific settings. First, when the complete-case estimator is a U-Statistic, we show that the optimal adjustment term also takes the form of a U-statistic. This motivates us to consider a CAM estimator with a general U-statistic as the adjustment term and allows for detailed theoretical analysis. In particular, when the data is missing completely at random, we show that our CAM estimator is unbiased, asymptotically Gaussian, and has a smaller asymptotic variance than the complete-case U-Statistic (cf. Theorem 3). We provide two concrete examples where a clear improvement can be shown analytically. Moreover, the numerical properties of the CAM estimator are demonstrated using simulated and real data, including an application using the *Terneuzen birth cohort* dataset available from `CRAN`.

We then investigate an application of the CAM technique to kernel based methods in nonparametric density estimation and regression problems. Under standard nonparametric assumptions on the data generating distribution, we quantify the leading order asymptotic improvement in mean squared error obtained by the CAM estimator compared with the complete-case approach. We provide further theoretical justification for the construction of the CAM estimator in these settings, in terms of approximating the optimal adjustment term. This leads to a fast, fully data-driven construction of our estimator. Finally, we demonstrate our method and the improvement it offers over the compete-case estimator in a simulation study and show how it can be used in an application with the Brandsma dataset on `CRAN`.

Related methods to the CAM approach have been utilised in various double-sample design settings. Fuller (1998) considers the problem of marginal mean estimation when one observes a small set of pairs of data and a larger set of univariate observations. Chen and Chen (2000) proposed an estimator of the regression parameters in a generalised linear model, where the practitioner has one sample in which the observations may be noisy or proxy versions of the variables of interest, and a second validation sample where complete and exact observations of the features are available. Chen and Chen (2000) show that their estimator of the regression parameter is asymptotically unbiased and has smaller (asymptotic) variance than a naive estimator based solely on the validation sample. Similar ideas have been used more recently in different statistical problems. Jiang et al. (2011) propose a nonparametric kernel-based regression estimate in double sampling designs, where the response is missing in one of the samples, but a surrogate outcome is observed instead. Yang and Ding (2020) propose an estimator of the average causal treatment effect in a general setting, by combining multiple observational datasets; see also Lin and Chen (2014), who focus on the logistic regression setting. Very recently, Zhang et al. (2019) considered estimating the marginal mean response in a semi-supervised setting, where one has a large number of *unlabelled* observations alongside a small labelled training dataset. Our work extends these existing works to a much wider range of estimation problems, including estimators based on U-Statistics, and nonparametric density estimation and regression.

Finally, we note here that, in contrast to many imputation approaches, the CAM technique has a clear and direct aim – that is to reduce the mean squared error of the complete-case estimator. On the other hand, while imputation is widely applicable, the properties of estimators derived from imputed data are typically not well understood. Indeed, imputation methods are often treated as a black-box solution to missing data problems, and the subsequent steps of estimation and inference are conducted independently of the imputation step, which may lead to unreliable results. In our numerical work in Section 4.3, we see that a regression estimator computed with imputed data can in fact perform worse than simply using a complete-case approach. Moreover, in problems such as density estimation, it is unclear whether imputation offers a viable solution – the distribution of the imputed data may not be the same as the target distribution.

The remainder of this paper is as follows. In Section 2 we fix our general statistical and missing data settings and introduce the CAM estimator. Section 3 is dedicated to studying U-Statistics. We then demonstrate how the CAM estimator can be applied using kernel methods in density estimation and regression problems in Section 4. We also provide, throughout the paper, a number of practical demonstrations of the method using

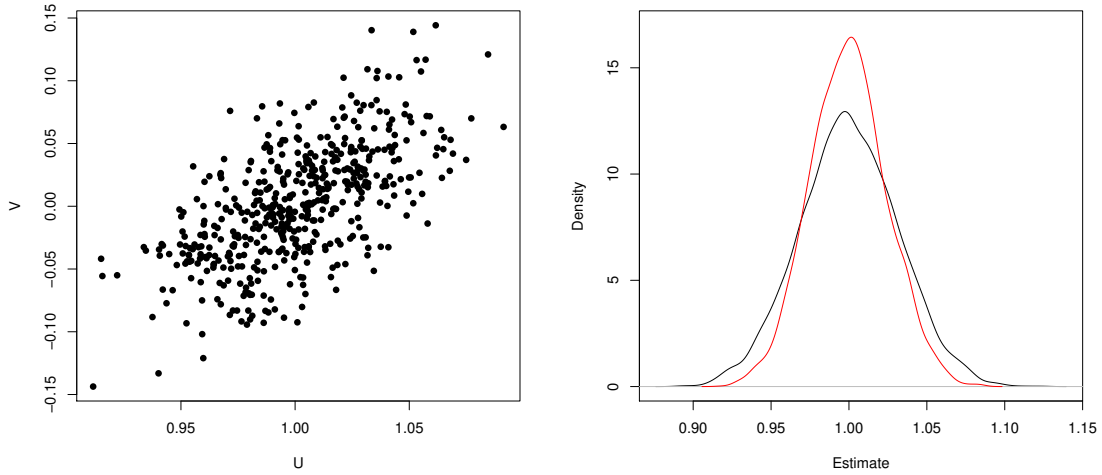

Figure 1: Left: 500 observations from the joint distribution of (U, V) (see (1)), with $\nu_X = \nu_Y = 1$, $\Gamma = \frac{1}{10}I + \frac{9}{10}(1,1)^T(1,1)$ and $n = 1000$. Right: Sampling distributions of $\hat{\nu}_{X,1}$ in black and $\tilde{\nu}_X$ in red.

real and simulated datasets. We conclude our paper with a discussion of various practical considerations and possible extensions in Section 5. All technical details and proofs of our theoretical results are presented in Section A in the appendix. We first end this section with an illustrative example that demonstrates how our estimator is constructed.

1.1 Illustrative example

Let $(X_1, Y_1)^T, \ldots, (X_{2n}, Y_{2n})^T$ be independent and identically distributed bivariate Gaussian random variables with unknown mean $\nu = (\nu_X, \nu_Y)^T \in \mathbb{R}^2$, but known covariance $\Gamma = (\Gamma_{ij})$; we write $N_2(\nu, \Gamma)$ as shorthand. Suppose we observe $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and $\{Y_{n+1}, \ldots, Y_{2n}\}$; in other words, in the language of missing data, the first component is missing completely at random in the second set of observations. Our task is to estimate ν_X.

The complete-case estimator in this example ignores the second set of observations and takes the sample mean of the X observations only. That is $\hat{\nu}_{X,1} := \frac{1}{n}\sum_{i=1}^n X_i \sim N(\nu_X, \Gamma_{11}/n)$; this is unbiased and is the maximum likelihood estimator if there are no Y observations. Now consider $\hat{\nu}_{Y,1} := \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\nu}_{Y,2} := \frac{1}{n}\sum_{i=n+1}^{2n} Y_i$. Then, letting $U := \hat{\nu}_{X,1}$ and $V := \hat{\nu}_{Y,1} - \hat{\nu}_{Y,2}$, we have

$$(U, V)^T \sim N_2((\nu_X, 0)^T, \tilde{\Gamma}), \tag{1}$$

where $\tilde{\Gamma} = n^{-1}\Gamma + n^{-1}\Gamma_{22}(0,1)^T(0,1)$; see Figure 1. This motivates the estimator

$$\tilde{\nu}_X := \hat{\nu}_{X,1} - \frac{\tilde{\Gamma}_{12}}{\tilde{\Gamma}_{22}}(\hat{\nu}_{Y,1} - \hat{\nu}_{Y,2}) \sim N\Big(\nu_X, \frac{1}{n}\Big(\Gamma_{11} - \frac{\tilde{\Gamma}_{12}^2}{\tilde{\Gamma}_{22}}\Big)\Big). \tag{2}$$

We see that $\tilde{\nu}_X$ is also unbiased for ν_X, but has a strictly smaller variance than $\hat{\nu}_{X,1}$ whenever $\Gamma_{12} \neq 0$.

There is a neat connection between our estimator and the Rao–Blackwell Theorem in this example. The statistic $T = (T_1, T_2)^T = (\hat{\nu}_{X,1} - \hat{\nu}_{Y,1}\Gamma_{12}/\Gamma_{22}, \hat{\nu}_{Y,1} + \hat{\nu}_{Y,2})^T$ is sufficient (the technical details are presented in Section A.1) for ν, and the *Rao–Blackwellised* version of $\hat{\nu}_{X,1}$ is $\mathbb{E}(\hat{\nu}_{X,1}|T) = \tilde{\nu}_X$. In fact, one can show that $\tilde{\nu}_X$ is the maximum likelihood estimator of ν_X in this setting (cf. Anderson (1957)).

2. Missing data and the CAM estimator

Suppose $Z = (X, Y)$ is a random pair taking values in $\mathbb{R}^d \times \mathbb{R}$ with joint distribution P. We are interested in estimating $\theta \equiv \theta(P) \in \mathbb{R}$, which may be any real valued function of the distribution P. Examples studied in detail in this paper include the mean of the first component of X; the covariance between X and Y; the value of a regression function of Y on X at the point $x \in \mathbb{R}^d$; or the value of the density $f_X(x)$, if it exists, of X at $x \in \mathbb{R}^d$.

We study a setting where some of the features of X are missing, but where the response Y is always observed. In order to model this, suppose that we have (Z, M), where the marginal distribution of Z is P, and M is a *missingness* indicator taking values in $\{0,1\}^d$. More precisely, we only observe the features $j \in \{1, \ldots, d\}$, with $M^j = 0$, that is $Z^M := (X^M, Y)$, where for $x = (x^1, \ldots, x^d)^T \in \mathbb{R}^d$ and $m = (m^1, \ldots, m^d)^T \in \{0,1\}^d$, we write $x^m := (x^j : m^j = 0) \in \mathbb{R}^{d_m}$, where $d_m := \sum_{j=1}^d \mathbb{1}_{\{m^j=0\}}$. Define the missingness probabilities $p_m(x, y) := \mathbb{P}(M = m | X = x, Y = y)$. Further write P_m and Q_m for the joint distribution of (X^m, Y) and $(X^m, Y)|\{M = m\}$, respectively. Note that under certain missing data assumptions we have $P_m = Q_m$, but that this is not the case in general.

We say the data is *missing completely at random* (MCAR) if M is independent of the pair (X, Y). In this case we write $p_m := \mathbb{P}(M = m)$. The data is *missing at random* (MAR) if the missingness indicator only depends on the observed data. Formally, this means that M is conditionally independent of Z given Z^M. Whereas, the data is said to be *missing not at random* (MNAR) if M depends on the unobserved value. See, for example, Little and Rubin (2002, Chapter 1.3) for further discussion of these three scenarios. Our most general results in this paper make no assumption on the missingness type. However, if the data is missing not at random, then the parameter of interest $\theta(P)$ may not be identifiable from the missing data and estimation will be problematic; see, for example, Wang et al. (2014) and Miao et al. (2016). For our technical analysis in the U-Statistics setting and the nonparametric learning problems, we focus on the MCAR case. Further discussion of the more challenging non-MCAR scenarios is given in Section 5.

Let $(Z_1, M_1), \ldots, (Z_n, M_n)$ be independent and identically distributed pairs with the same distribution as (Z, M). It is convenient to consider the missingness indicators $M_1, \ldots, M_n$ as fixed and equal to $m_1, \ldots, m_n$, and from this point on all probability statements should be interpreted to be conditional on $(M_1, \ldots, M_n) = (m_1, \ldots, m_n)$. We observe $Z_1^{m_1}, \ldots, Z_n^{m_n}$. The popular complete-case approach uses only the observations with $m_i = (0, \ldots, 0) \in \mathbb{R}^d$. Our goal in this paper is to construct an estimator which also uses the observations with missing values in order to improve on the performance of the complete-case estimator.

It is also useful to introduce some further notation here. For $m \in \{0,1\}^d$, let $A_m := \{i \in \{1,\dots,n\} : m_i = m\}$ be the set of indices of the data missing m, let $n_m := |A_m|$. In particular, A_0 is the set of indices of the complete cases, where here and throughout we use the shorthand 0 in place of $0_d := (0,\dots,0) \in \mathbb{R}^d$. We have that $A_{m_1} \cap A_{m_2} = \emptyset$ for $m_1 \neq m_2$. Finally, for a set $A \subseteq \{1,\dots,n\}$ and $m \in \{0,1\}^d$, let $\mathcal{T}_{A,m} := \{Z_i^m : i \in A\}$. Of course, $\mathcal{T}_{A,m}$ is not necessarily observed for every A and m, but we do observe $\mathcal{T}_{A_m,m}$. We assume that A_0 is non-empty, and moreover, for each $m \in \{0,1\}^d$ with A_m non-empty, we have that $\lim_{n\to\infty} \frac{n_m}{n} = q_m \in (0,1)$, almost surely. That is, either a set of features m is never missing, or, for those that are missing, the relative sample sizes are more or less balanced. If the data is MCAR, then we have that $q_m = p_m$, for $m \in \{0,1\}^d$.

We now define a generic version of our *correlation-assisted missing data* (CAM) estimator. The main idea underpinning our proposal is to mimic the approach in the toy example in Section 1.1 by combining appropriate, correlated estimators, which are constructed using different parts of the data. We first assume that there is a suitable complete-case estimator of θ that only uses the data in $\mathcal{T}_{A_0,0}$; this is denoted by $\hat{\theta}_0 = \hat{\theta}_{A_0,0}$. Furthermore, suppose that, for each m and each $A \subseteq \{1,\dots,n\}$, we have access to a statistic denoted by $\hat{\varphi}_{A,m}$, which only depends on the data in $\mathcal{T}_{A,m}$. Notice that, for $m \neq 0$, $\hat{\varphi}_{A,m}$ is not an estimator for θ. A detailed discussion of the choice of $\hat{\varphi}_{A,m}$ is given at the end of this section.

Consider $m \in \{0,1\}^d \setminus \{0_d\}$ such that A_m is non-empty. The CAM estimator is constructed using $\hat{\theta}_0$, $\hat{\varphi}_{A_m,m}$ and $\hat{\varphi}_{A_0,m}$. The hope is that the latter two statistics have similar expected values, and that $\hat{\varphi}_{A_0,m}$ is (highly) correlated with the complete-case estimator $\hat{\theta}_0$. We then exploit this correlation in the same way as we did in the illustrative example – see (2).

In fact, we can construct similar statistics using many $m \in \{0,1\}^d \setminus \{0_d\}$ simultaneously. Let $\mathcal{M} \subseteq \{0,1\}^d \setminus \{0_d\}$ denote the set of values of m that we would like to use (a detailed discussion of how to choose $\mathcal{M}$ in practice is postponed until Section 5). Consider the two column vectors of length $|\mathcal{M}|$, given by $\hat{\varphi}_{0,\mathcal{M}} := \big(\{\hat{\varphi}_{A_0,m} : m \in \mathcal{M}\}\big)^T$ and $\hat{\varphi}_{\mathcal{M}} := \big(\{\hat{\varphi}_{A_m,m} : m \in \mathcal{M}\}\big)^T$. Finally, for $\gamma \in \mathbb{R}^{|\mathcal{M}|}$, define the *correlation-assisted missing data* (CAM) estimator

$$\hat{\theta}_\gamma^{\mathcal{M}} := \hat{\theta}_0 - \gamma^T(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}).$$

In practice we will use a data-driven choice of γ, which aims to minimise the mean squared error (cf. the discussion after the statement of Proposition 1). It's perhaps useful to relate back to the illustrative example in Section 1.1: there we have $\mathcal{M} = \{m\} = \{1\}$, $\hat{\theta}_0 = \hat{\nu}_{X,1}$, $\hat{\varphi}_{0,m} = \hat{\nu}_{Y,1}$, $\hat{\varphi}_m = \hat{\nu}_{Y,2}$ and $\gamma = \tilde{\Gamma}_{12}/\tilde{\Gamma}_{22}$ (note also that $A_0 = \{1,\dots,n\}$ and $A_m = \{n+1,\dots,2n\}$).

We now study properties of the CAM approach in the general estimation problem. For an estimator $\hat{\theta}$, let $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\{(\hat{\theta}-\theta)^2\}$ denote its mean squared error. Let $b(\hat{\theta}) := \mathbb{E}(\hat{\theta}-\theta)$ be the bias of an estimator $\hat{\theta}$ and let $B_{\mathcal{M}} := \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})$. Let Ω denote the $|\mathcal{M}|$-dimensional vector of covariances $\mathrm{Cov}(\hat{\theta}_0, \hat{\varphi}_{0,\mathcal{M}})$ and let Λ be the $|\mathcal{M}| \times |\mathcal{M}|$ covariance matrix $\mathrm{Var}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})$.

Proposition 1 *We have that*

$$\mathrm{MSE}(\hat{\theta}_\gamma^{\mathcal{M}}) - \mathrm{MSE}(\hat{\theta}_0) = \gamma^T(\Lambda + B_{\mathcal{M}}B_{\mathcal{M}}^T)\gamma - 2\gamma^T\{\Omega + b(\hat{\theta}_0)B_{\mathcal{M}}\}.$$

In particular, if Λ is nonsingular and $B_{\mathcal{M}} = 0$, then $\gamma = \gamma^* := \Lambda^{-1}\Omega$ is the optimal weight vector achieving the maximum reduction in MSE, and

$$\mathrm{MSE}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) - \mathrm{MSE}(\hat{\theta}_0) = -\Omega^T\Lambda^{-1}\Omega \leq 0.$$

Proposition 1 compares the MSE of our CAM estimator and the complete-case estimator. We have made no assumption on the missing data mechanism and, in particular, we do not assume here that the data is missing completely at random. It is worth noting, however, that if the data is not MCAR, the complete-case estimator may be (even asymptotically) biased (cf. Section 5). In which case, simply improving on the performance of the complete-case estimator will not necessarily be effective. Furthermore, for general missing data mechanisms, we do not have control of $B_{\mathcal{M}}$. However, we will see that under appropriate conditions in many estimation problems, $\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}$ will be (asymptotically) mean zero.

The role of $\hat{\varphi}_m$ here is to estimate $\mathbb{E}(\hat{\varphi}_{0,m})$. If we can reduce the MSE of this estimator, then this may lead in turn to an overall reduction in MSE for estimating θ. To see this, suppose $\mathcal{M} = \{m\}$ and $b(\hat{\theta}_0) = 0$, then the best MSE reduction for estimating θ is $\frac{\mathrm{Cov}^2(\hat{\theta}_0,\hat{\varphi}_{0,m})}{\mathrm{Var}(\hat{\varphi}_{0,m})+\mathrm{MSE}(\hat{\varphi}_m)}$, where $\mathrm{MSE}(\hat{\varphi}_m) = \mathbb{E}\{(\hat{\varphi}_m - \mathbb{E}\hat{\varphi}_{0,m})^2\} = \mathrm{Var}(\hat{\varphi}_m) + B^2_{\{m\}}$. We see, then, that one should choose $\hat{\varphi}_m$ to minimise the MSE in estimating $\mathbb{E}(\hat{\varphi}_{0,m})$. On the other hand, if the complete case-estimator is biased, then the effect of minimising the MSE of $\hat{\varphi}_m$ is unclear.

We see from the second part of Proposition 1 that to achieve maximum mean squared error reduction when $B_{\mathcal{M}} = 0$, we should set $\gamma = \gamma^* := \Lambda^{-1}\Omega$. If, moreover, $\mathcal{M} = \{m\}$, then we have that

$$\mathrm{MSE}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) - \mathrm{MSE}(\hat{\theta}_0) = -\mathrm{Var}(\hat{\theta}_0)\frac{\mathrm{Var}(\hat{\varphi}_{0,m})}{\mathrm{Var}(\hat{\varphi}_{0,m}) + \mathrm{Var}(\hat{\varphi}_m)}\mathrm{Corr}^2(\hat{\theta}_0, \hat{\varphi}_{0,m}), \tag{3}$$

where $\mathrm{Corr}(\hat{\theta}_0, \hat{\varphi}_{0,m}) = \frac{\mathrm{Cov}(\hat{\theta}_0,\hat{\varphi}_{0,m})}{\sqrt{\mathrm{Var}(\hat{\theta}_0)\mathrm{Var}(\hat{\varphi}_{0,m})}}$. Here we have used that $\hat{\varphi}_{0,m}$ and $\hat{\varphi}_m$ are independent since they are constructed using disjoint sets of observations. Thus, to achieve a maximal reduction in MSE, we'd like $\hat{\varphi}_{0,m}$ to be maximally correlated with $\hat{\theta}_0$ and $\mathrm{Var}(\hat{\varphi}_m)$ to be minimised. The first is achieved by the conditional expectation $\hat{\varphi}^*_{0,m} = \mathbb{E}(\hat{\theta}_0|\mathcal{T}_{A_0,m})$. Moreover, $\mathrm{Var}(\hat{\varphi}_m)$ is minimised by $\hat{\varphi}^*_m = \mathbb{E}(\hat{\varphi}_{0,m})$, but this is typically unknown. In practice, we use the data $\mathcal{T}_{A_m,m}$ to construct an estimate of $\mathbb{E}(\hat{\varphi}_{0,m})$ that has low variance. The situation when $|\mathcal{M}| > 1$ is similar by noting the independence of $\mathcal{T}_{A_{m_1},m_1}$ and $\mathcal{T}_{A_{m_2},m_2}$, for $m_1 \neq m_2$.

To understand this further, note that with the optimal choice $\hat{\varphi}^*_{0,m}$ we have

$$\mathrm{Cov}(\hat{\theta}_0, \hat{\varphi}^*_{0,m}) = \mathbb{E}\{\mathrm{Cov}(\hat{\theta}_0, \hat{\varphi}^*_{0,m}|\mathcal{T}_{A_0,m})\} + \mathrm{Cov}\{\mathbb{E}(\hat{\theta}_0|\mathcal{T}_{A_0,m}), \mathbb{E}(\hat{\varphi}^*_{0,m}|\mathcal{T}_{A_0,m})\} = \mathrm{Var}(\hat{\varphi}^*_{0,m}).$$

Thus, in the ideal case that $\mathrm{Var}(\hat{\varphi}_m)$ is negligible compared with $\mathrm{Var}(\hat{\varphi}^*_{0,m})$ (e.g. if $n_m \gg n_0$), the improvement in MSE is simply $\mathrm{Var}(\hat{\varphi}^*_{0,m}) = \mathrm{Var}\{\mathbb{E}(\hat{\theta}_0|\mathcal{T}_{A_0,m})\}$.

Of course, the conditional expectation $\mathbb{E}(\hat{\theta}_0|\mathcal{T}_{A_0,m})$ is also typically unknown. We will see in practice that, for instance, assuming a parametric form for $\hat{\varphi}^*_{0,m}$ works well. In particular, in our study of U-Statistics in Section 3, we see that a data-driven choice of $\hat{\varphi}_{0,m}$ will often lead to similar performance to the optimal choice. Moreover, for nonparametric methods using kernels, the optimal $\hat{\varphi}^*_{0,m}$ can often be well approximated by the same type of nonparametric estimator with a practical choice of kernel (see Section 4).

3. U-Statistics

In this section we specialise to the setting of U-Statistics. Suppose we are interested in estimating a parameter of the form $\theta = \theta(P) = \mathbb{E}\{\phi(Z_1, \ldots, Z_r)\}$, for $r \geq 1$ and some function $\phi : (\mathbb{R}^d \times \mathbb{R})^{\otimes r} \to \mathbb{R}$, which is permutation symmetric in its r arguments (i.e. we have that $\phi(Z_1, \ldots, Z_r) = \phi(Z_{\sigma(1)}, \ldots, Z_{\sigma(r)})$, for all permutations $\sigma : \{1, \ldots, r\} \to \{1, \ldots, r\}$). In the non-missing setting, an unbiased estimator of θ is given by

$$\hat{\theta} = \frac{1}{\binom{n}{r}} \sum_{\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}} \phi(Z_{i_1}, \ldots, Z_{i_r}),$$

where the sum is taken over all unordered subsets $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}$ of size r. Statistics of this form have been studied in detail in the non-missing setting, see for instance van der Vaart (1998, Chapter 12.1). In particular, if $\mathbb{E}\{\phi^2(Z_1, \ldots, Z_r)\} < \infty$, then $n^{1/2}(\hat{\theta} - \theta) \to^d N(0, r^2\psi_U)$, where $\psi_U = \mathrm{Cov}\{\phi(Z_1, \ldots, Z_r), \phi(Z_1, Z_{r+1}, \ldots, Z_{2r-1})\}$.

We now construct the CAM U-Statistic. First, the complete-case U-Statistic is

$$\hat{\theta}_0 = \hat{\theta}_{A_0,0} = \frac{1}{\binom{n_0}{r}} \sum_{\{i_1, \ldots, i_r\} \subseteq A_0} \phi(Z_{i_1}, \ldots, Z_{i_r}),$$

where now the sum is taken over all unordered subsets $\{i_1, \ldots, i_r\} \subseteq A_0$ of size r. In this case, for $m \in \mathcal{M}$, we have

$$\mathbb{E}(\hat{\theta}_0 | \mathcal{T}_{A_0,m}) = \frac{1}{\binom{n_0}{r}} \sum_{\{i_1, \ldots, i_r\} \subseteq A_0} \mathbb{E}\{\phi(Z_{i_1}, \ldots, Z_{i_r}) | Z_{i_1}^m, \ldots, Z_{i_r}^m\}.$$

In other words, the optimal form of the adjustment term in the construction of the CAM estimator is itself a U-Statistic with oracle kernel

$$\phi_m^*(Z_1^m, \ldots, Z_r^m) = \mathbb{E}\{\phi(Z_1, \ldots, Z_r) | Z_1^m, \ldots, Z_r^m\}. \tag{4}$$

This depends on ϕ and the conditional distribution of Z given Z^m, which is typically unknown. We therefore consider a general construction of the adjustment term as follows: for $m \in \mathcal{M}$, let $\phi_m : (\mathbb{R}^{d_m} \times \mathbb{R})^{\otimes r} \to \mathbb{R}$ be a permutation symmetric function in its r arguments, and, for $A \subseteq \{1, \ldots, n\}$, define

$$\hat{\varphi}_{A,m} = \frac{1}{\binom{|A|}{r}} \sum_{\{i_1, \ldots, i_r\} \subseteq A} \phi_m(Z_{i_1}^m, \ldots, Z_{i_r}^m).$$

We will make use of $\hat{\varphi}_{0,m} = \hat{\varphi}_{A_0,m}$ and $\hat{\varphi}_m = \hat{\varphi}_{A_m,m}$. Recall that $\hat{\varphi}_{0,\mathcal{M}} = (\hat{\varphi}_{0,m} : m \in \mathcal{M})^T$ and $\hat{\varphi}_{\mathcal{M}} = (\hat{\varphi}_m : m \in \mathcal{M})^T$. For $\gamma \in \mathbb{R}^{|\mathcal{M}|}$, define the CAM U-Statistic

$$\hat{\theta}_\gamma^{\mathcal{M}} := \hat{\theta}_0 - \gamma^T(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}).$$

Here ϕ_m is left unspecified, in practice one would aim to choose ϕ_m to mimic the oracle choice above.

Suppose that the data is missing completely at random. Then we have that $B_{\mathcal{M}} = \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}) = 0$. Recall also that $\Omega = \mathrm{Cov}(\hat{\theta}_0, \hat{\varphi}_{0,\mathcal{M}})$ and $\Lambda = \mathrm{Var}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})$. It follows directly from Proposition 1 that $\mathrm{MSE}(\hat{\theta}_\gamma^{\mathcal{M}}) - \mathrm{MSE}(\hat{\theta}_0) = \gamma^T \Lambda \gamma - 2\gamma^T \Omega$.

Our next two results concern the asymptotic properties of the CAM U-Statistic. Let Ω_U be the $|\mathcal{M}|$-dimensional vector with entries

$$\Omega_{U,m} := \mathrm{Cov}\big\{\phi\big(Z_1, \ldots, Z_r\big), \phi_m\big(Z_1^m, Z_{r+1}^m, \ldots, Z_{2r-1}^m\big)\big\}.$$

Further, let Λ_U be the $|\mathcal{M}| \times |\mathcal{M}|$ symmetric matrix with diagonal entries

$$\Lambda_{U,m,m} := \Big(1 + \frac{p_0}{p_m}\Big)\mathrm{Cov}\big\{\phi_m\big(Z_1^m, \ldots, Z_r^m\big), \phi_m\big(Z_1^m, Z_{r+1}^m, \ldots, Z_{2r-1}^m\big)\big\}$$

and off-diagonal entries

$$\Lambda_{U,m_1,m_2} := \mathrm{Cov}\big\{\phi_{m_1}\big(Z_1^{m_1}, \ldots, Z_r^{m_1}\big), \phi_{m_2}\big(Z_1^{m_2}, Z_{r+1}^{m_2}, \ldots, Z_{2r-1}^{m_2}\big)\big\}.$$

We see in Theorem 2 that, under moment assumptions, we have $\Omega \to \Omega_U$ and $\Lambda \to \Lambda_U$ as $n \to \infty$, and that the CAM U-Statistic is unbiased and asymptotically Gaussian.

Theorem 2 *Suppose the data is missing completely at random,* $\mathbb{E}\{\phi^2\big(Z_1, \ldots, Z_r\big)\} < \infty$ *and, for* $m \in \mathcal{M}$, $\mathbb{E}\{\phi_m^2\big(Z_1^m, \ldots, Z_r^m\big)\} < \infty$. *Then, for* $\gamma \in \mathbb{R}^{|\mathcal{M}|}$,

$$\sqrt{n_0}(\hat{\theta}_\gamma^{\mathcal{M}} - \theta) \to^d N\big(0, r^2(\psi_U + \gamma^T \Lambda_U \gamma - 2\gamma^T \Omega_U)\big)$$

as $n \to \infty$.

Theorem 2 shows that an asymptotically optimal choice of γ, which minimises the asymptotic variance of the CAM U-Statistic, is $\gamma^* = \Lambda_U^{-1}\Omega_U$. The optimal leading order asymptotic variance reduction is $n_0^{-1} r^2 \Omega_U^T \Lambda_U^{-1} \Omega_U$. This is of the same order as the asymptotic variance of the complete-case U-Statistic, which is $n_0^{-1} r^2 \psi_U$. Of course Ω_U and Λ_U are typically unknown. However, to estimate these we can further exploit the use of U-Statistics. Note that

$$\Omega_{U,m} = \frac{1}{2}\mathbb{E}\big[\big\{\phi(Z_1, \ldots, Z_r) - \phi(Z_{2r}, \ldots, Z_{3r-1})\big\} \\ \big\{\phi_m(Z_1^m, Z_{r+1}^m, \ldots, Z_{2r-1}^m) - \phi_m(Z_{2r}^m, Z_{3r}^m, \ldots, Z_{4r-2}^m)\big\}\big].$$

Thus, we can estimate Ω_U using a U-Statistic of order $4r - 2$ (see (16) in Section A.3). A similar expression can be derived for the entries of Λ_U, but for brevity we exclude the formulas here – they are given in (17) and (18) in Section A.3. Let $\hat{\Lambda}_U$ and $\hat{\Omega}_U$ denote resulting U-Statistic estimators of Λ_U and Ω_U, respectively. (In practice, averaging over all subsamples of size $4r - 2$ will be computationally expensive, in our simulations $\hat{\Omega}_U$ and $\hat{\Lambda}_U$ are approximated using 10^5 random subsamples.)

Now, let $\hat{\gamma} := \hat{\Lambda}_U^{-1}\hat{\Omega}_U$ and consider the practical CAM U-Statistic $\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}}$. Theorem 3 shows that we can mimic the performance of the optimal CAM U-Statistic $\hat{\theta}_{\gamma^*}^{\mathcal{M}}$ using the data-driven choice of γ.

Theorem 3 *Suppose the data is missing completely at random,* $\mathbb{E}\{\phi^4(Z_1,\ldots,Z_r)\} < \infty$ *and, for* $m \in \mathcal{M}$, $\mathbb{E}\{\phi_m^4(Z_1^m,\ldots,Z_r^m)\} < \infty$. *Then*

$$\sqrt{n_0}(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \theta) \to^d N\big(0, r^2(\psi_U - \Omega_U^T \Lambda_U^{-1} \Omega_U)\big). \tag{5}$$

We remark that in fact (5) holds with $\hat{\gamma}$ replaced by any other consistent estimator of γ^*. In particular, if we estimate Ω_U and Λ_U using incomplete U-Statistics based on $B = B_{n_0}$ randomly chosen subsamples, then (5) holds so long as $B_{n_0} \to \infty$ as $n_0 \to \infty$. As mentioned above, in our numerical experiments we set $B = 10^5$. Moreover if $\hat{\theta}_0$, $\hat{\varphi}_{0,m}$ and $\hat{\varphi}_m$ are themselves replaced by the incomplete U-statistics based on B randomly chosen subsamples, then the conclusion in (5) holds so long as $B^{-1}\max_m\{n_m\} \to 0$; see, for example, Janson (1984, Corollary 1). One crucial aspect here would be to ensure that the same subsamples are used in the construction of $\hat{\theta}_0$ and $\hat{\varphi}_{0,m}$ in order to ensure that the covariance between these two estimators is as large as possible.

The asymptotic variance in (5) can be estimated by plugging in the estimators of ψ_U, Ω_U, and Λ_U. Here a U-Statistic estimator of ψ_U, denoted by $\hat{\psi}_U$, can be constructed in the same way as $\hat{\Omega}_U$ and $\hat{\Lambda}_U$. Then one can show that $\sqrt{n_0}(\hat{\psi}_U - \hat{\Omega}_U^T \hat{\Lambda}_U^{-1} \hat{\Omega}_U)^{-1/2}(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \theta) \to^d N(0, r^2)$, which can be used for statistical inference such as constructing confidence intervals and testing hypotheses. For example, an asymptotic $(1-\alpha)100\%$ confidence interval can be constructed as

$$\left(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \frac{r z_{\alpha/2}}{\sqrt{n_0}}\sqrt{\hat{\psi}_U - \hat{\Omega}_U^T \hat{\Lambda}_U^{-1} \hat{\Omega}_U}, \quad \hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} + \frac{r z_{\alpha/2}}{\sqrt{n_0}}\sqrt{\hat{\psi}_U - \hat{\Omega}_U^T \hat{\Lambda}_U^{-1} \hat{\Omega}_U}\right),$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ denotes the standard normal distribution function.

Remark 4 *The form of oracle kernel in* (4) *suggests the possibility of using a data driven kernel constructed via data splitting. Specifically, suppose we have access to another complete case dataset* $\tilde{A}_0$ *with* $\tilde{n} := |\tilde{A}_0|$, *and we use some mean regression technique, either parametric or nonparametric, to obtain an estimate* $\tilde{\phi}_m$ *of* ϕ_m^* *satisfying* $\|\tilde{\phi}_m - \phi_m^*\|_\infty = O_p(a_{\tilde{n}})$ *with* $a_{\tilde{n}} \to 0$ *as* $\tilde{n} \to \infty$. *Denote by* $\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}$ *and* $\tilde{\varphi}_{0,\mathcal{M}} - \tilde{\varphi}_{\mathcal{M}}$ *the bias adjustment term using the oracle kernel and data driven kernel, respectively. Then we have*

$$\|(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}) - (\tilde{\varphi}_{0,\mathcal{M}} - \tilde{\varphi}_{\mathcal{M}})\|_2 = O_p(a_{\tilde{n}}),$$

as $\tilde{n} \to \infty$. *Consequently, for* $\tilde{\theta}_\gamma^{\mathcal{M}} := \hat{\theta}_0 - \gamma^T(\tilde{\varphi}_{0,\mathcal{M}} - \tilde{\varphi}_{\mathcal{M}})$, *we have*

$$|\tilde{\theta}_\gamma^{\mathcal{M}} - \hat{\theta}_\gamma^{\mathcal{M}}| \le \|\gamma\|_2 \cdot \|(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}) - (\tilde{\varphi}_{0,\mathcal{M}} - \tilde{\varphi}_{\mathcal{M}})\|_2 = \|\gamma\|_2 O_p(a_{\tilde{n}}).$$

Thus, the asymptotic normality in Theorem 2 holds if $a_{\tilde{n}}\sqrt{n_0} \to 0$ *and* $\|\gamma\|_2$ *is bounded. The condition* $a_{\tilde{n}}\sqrt{n_0} \to 0$ *implicitly requires that* $\tilde{n}/n_0 \to \infty$, *suggesting that a large independent complete case dataset is needed to make the estimation error of the kernel negligible, in which case it is likely that complete case estimator based on* $A_0 \cup \tilde{A}_0$ *would perform better. We therefore choose not to pursue this data-splitting idea further.*

In order to understand the improvement the CAM U-Statistic achieves over the complete-case method it is helpful to consider some examples.

Example 1 (Marginal mean estimation) *Suppose we are interested in estimating the parameter $\theta = \mathbb{E}(X)$ of the pair $(X, Y) \in \mathbb{R}^2$ (i.e. $d = 1$). Suppose further that the X variable is missing completely at random with probability 0.5. Here, then, m_i is either 0 (if x_i is not missing) or 1 (if x_i is missing), and $A_0 = \{i \in \{1, \ldots, n\} : m_i = 0\}$ and $A_1 = \{i \in \{1, \ldots, n\} : m_i = 1\}$, where the respective sample sizes are n_0 and n_1, and $n = n_0 + n_1$. In contrast to the illustrative example in Section 1.1, we are not assuming the joint distribution of (X, Y) is Gaussian. In this setting the complete-case U-Statistic is $\hat{\theta}_0 = \frac{1}{n_0}\sum_{i\in A_0} X_i$. Of course, by the Central Limit Theorem, if $\mathbb{E}(X^2) < \infty$, then $\sqrt{n_0}(\hat{\theta}_0 - \theta) \to^d N(0, \mathrm{Var}(X))$.*

Next we consider the CAM estimator with $\mathcal{M} = \{1\}$ which takes into account the information from variable Y. For a generic function $\phi_1 : \mathbb{R} \to \mathbb{R}$ satisfying $\mathrm{Var}(\phi_1(Y_1)) < \infty$ define

$$\hat{\varphi}_{0,1} = \frac{1}{n_0}\sum_{i\in A_0} \phi_1(Y_i); \quad \hat{\varphi}_1 = \frac{1}{n_1}\sum_{i\in A_1} \phi_1(Y_i).$$

Then $\Lambda = \big(\frac{1}{n_0} + \frac{1}{n_1}\big)\mathrm{Var}\{\phi_1(Y_1)\}$ and $\Omega = \frac{1}{n_0}\mathrm{Cov}\{X_1, \phi_1(Y_1)\}$. We have that

$$\mathrm{MSE}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) - \mathrm{MSE}(\hat{\theta}_0) = \mathrm{Var}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) - \mathrm{Var}(\hat{\theta}_0) = -\frac{n_1 \mathrm{Cov}^2\{X, \phi_1(Y)\}}{n_0 n \mathrm{Var}\{\phi_1(Y)\}} \leq 0.$$

Thus, there is a guaranteed improvement in MSE as long as X and $\phi_1(Y)$ are correlated.

The optimal choice of ϕ_1 in this case is $\phi_1^*(y) := \mathbb{E}(X|Y = y)$, (cf. the discussion at the end of the previous section), and the corresponding first order variance reduction is $\frac{n_1}{n_0 n}\mathrm{Var}\{\mathbb{E}(X|Y)\}$. If X and Y are independent, then $\mathrm{Var}\{\mathbb{E}(X|Y)\} = 0$, i.e. as expected, the CAM estimator will not lead to an improvement over the complete-case estimator, since the Y variable tells us nothing about the marginal X distribution. On the other hand, in the pathological case that X can be written as a deterministic function of Y, we see that the variance reduction is $\frac{n_1}{n_0 n}\mathrm{Var}(X)$. Consequently, $\mathrm{Var}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) = \frac{\mathrm{Var}(X)}{n_0}\big(1 - \frac{n_1}{n}\big) = \frac{\mathrm{Var}(X)}{n}$, i.e. the variance that could be achieved by using a fully observed dataset!

Of course the regression function $\mathbb{E}(X|Y = y)$ will typically be unknown to the user. Consider instead therefore the practical choice $\phi_m(y) := y$. Then we have that

$$\mathrm{Var}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) = \frac{\mathrm{Var}(X)}{n_0}\Big\{1 - \frac{n_1}{n}\mathrm{Corr}^2(X, Y)\Big\},$$

where $\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$ denotes the correlation between X and Y. In fact, the above derivation holds as long as $\phi_m(y)$ is a linear function of y.

In the left panel of Figure 2, we present the sampling distributions of the complete-case and CAM U-Statistic for the mean of $X \sim \mathrm{Exp}(1)$, where $Y|\{X = x\} \sim N(x, \sigma^2)$. We set $n = 1000$, $\sigma = 0.2$ and the X variable is missing with probability 0.5. We present the results for the practical choice $\phi_m(y) = y$ and the optimal choice $\phi_m(y) = \mathbb{E}(X|Y = y) = y - \sigma^2 + \sigma\frac{\Phi'(\sigma - y/\sigma)}{1-\Phi(\sigma - y/\sigma)}$, where $\Phi'(\cdot)$ and $\Phi(\cdot)$ denote the standard Normal density and distribution function, respectively. The variance reduction can be clearly seen from the plots. We see also that the practical CAM U-Statistic has very similar performance to the optimal CAM U-Statistic.

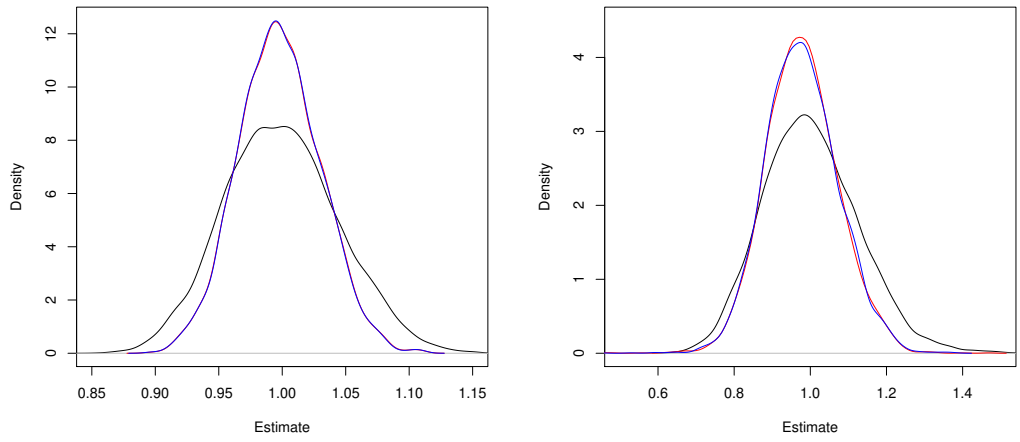

Figure 2: Sampling distributions of the complete-case U-Statistic $\hat{\theta}_0$ in black, the CAM U-Statistic $\hat{\theta}^{\mathcal{M}}_{\hat{\gamma}}$ in red and the optimal CAM U-Statistic (i.e. with $\phi_m = \phi^*_m$ and $\gamma = \gamma^*$) in blue for Example 1 (left) and 2 (right).

Example 2 (Covariance estimation) *Consider the same set-up as in Example 1, but suppose now we are interested in the parameter* $\theta = \mathrm{Cov}(X, Y) = \frac{1}{2}\mathbb{E}\{(X_1 - X_2)(Y_1 - Y_2)\}$. *In this case, we have the complete-case U-Statistic*

$$\hat{\theta}_0 = \frac{1}{2\binom{n_0}{2}} \sum_{\{i,j\} \subseteq A_0} (X_i - X_j)(Y_i - Y_j).$$

If $\mathbb{E}\{(X_1 - X_2)^2(Y_1 - Y_2)^2\} < \infty$, *then by van der Vaart (1998, Theorem 12.3) we have that* $\sqrt{n_0}(\hat{\theta}_0 - \theta) \to^d N(0, \psi_1)$, *where* $\psi_1 := \frac{1}{4}\mathrm{Cov}\{(X_1 - X_2)(Y_1 - Y_2), (X_1 - X_3)(Y_1 - Y_3)\}$.

Now, for a generic function $\phi_1 : \mathbb{R}^2 \to \mathbb{R}$ *consider*

$$\hat{\varphi}_{0,1} = \frac{2}{n_0(n_0 - 1)} \sum_{\{i,j\} \subseteq A_0} \phi_1(Y_i, Y_j); \quad \hat{\varphi}_1 = \frac{2}{n_1(n_1 - 1)} \sum_{\{i,j\} \subseteq A_1} \phi_1(Y_i, Y_j).$$

Here the optimal function is $\phi^*_1(y_1, y_2) = \frac{1}{2}\{\mathbb{E}(X|Y = y_1) - \mathbb{E}(X|Y = y_2)\}(y_1 - y_2)$. *Recall that* $p_1 = \lim_{n\to\infty} \frac{n_1}{n}$. *The corresponding CAM U-statistic satisfies*

$$\lim_{n\to\infty} n_0 \mathrm{Var}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) = \psi_1 - \frac{p_1}{4}\mathrm{Cov}\{\phi^*_1(Y_1, Y_2), \phi^*_1(Y_1, Y_3)\}.$$

Thus, we have first order variance reduction as long as $\mathrm{Cov}\{\phi^*_1(Y_1, Y_2), \phi^*_1(Y_1, Y_3)\} \neq 0$.

However, since ϕ^*_1 *is generally unknown, consider the practical choice* $\phi_1(y_1, y_2) = \frac{1}{2}(y_1 - y_2)^2$. *This is motivated by supposing that* $\mathbb{E}(X|Y = y)$ *is a linear function of* y. *Then we have*

$$\lim_{n\to\infty} n_0 \mathrm{Var}(\hat{\theta}^{\mathcal{M}}_{\gamma^*}) = \psi_1 - \frac{p_1 \mathrm{Cov}^2\{(X_1 - X_2)(Y_1 - Y_2), (Y_1 - Y_3)^2\}}{4\mathrm{Cov}\{(Y_1 - Y_2)^2, (Y_1 - Y_3)^2\}}$$

In the right panel of Figure 2, we present the sampling distributions of the complete-case, practical CAM, and optimal CAM U-Statistics for $\theta = \mathrm{Cov}(X, Y)$. As in Example 1, the data generating distribution is $X \sim \mathrm{Exp}(1)$ and $Y|\{X = x\} \sim N(x, \sigma^2)$, $n = 1000$, and $p_1 = 0.5$.

3.1 The Terneuzen birth cohort dataset

We now demonstrate how the CAM U-Statistic can be used in practice. In particular, we will apply our proposal from the previous two examples to the Terneuzen birth cohort data available from the `mice` package on CRAN. The full dataset consists of 3951 observations of 11 features covering 306 people. We simplify the problem by taking a subset of the data and only include the first measurement for each person. Furthermore, we retain only 4 of the features, namely "sex", "height Z-score", "weight Z-score", and "bmi Z-score". In the resulting dataset, there are 306 observations (one for each patient), of which 105 are missing both the height and bmi features. In order to fit this in the framework introduced above let Y denote sex (1 for female, 0 for male), and let X be the 3-dimensional vector of weight, height, and bmi.

We have 201 complete cases in A_0, and 105 cases in A_m for $m = (1, 0, 1)^T$, where only Y (sex) and $X^{(2)}$ (weight) are observed. We consider two problems; (i) to estimate the average bmi Z-score in the cohort, and (ii) estimate the covariance between the height and weight Z-scores. In both cases, we have $\mathcal{M} = \{(1, 0, 1)^T\}$, and we consider two choices of ϕ_m, for $m = (1, 0, 1)^T$, a simple choice, and a regression estimate.

What is X? also Am denotes?

In problem (i), recall that we can write the marginal mean as a U-Statistic with $\phi(Z) = X^{(3)}$, and the complete-case estimator is $\frac{1}{n_0}\sum_{i \in A_0} X_i^{(3)}$. To construct the CAM U-Statistics, we first consider $\phi_m(Z^m) = X^{(2)}$, i.e. the weight Z-score. For our second choice, write

$$\phi_m(Z^m) = \beta_0 + \beta_1 Y + \beta_2 X^{(2)} + \beta_3 X^{(2)} Y. \tag{6}$$

We choose $(\beta_0, \beta_1, \beta_2, \beta_3)$ by fitting a linear model (with an interaction) of bmi on height and sex using the complete cases (in `R` this is simply done using the `lm` function). The idea here is to approximate $\phi_m^*(Z^m) = \mathbb{E}\{\phi(Z)|Z^m\}$. Then $\hat{\varphi}_{0,m}$ is the sample average of the fitted values, and $\hat{\varphi}_m$ is the average of the predictions made on the data in A_m.

For problem (ii), we have $\phi(Z_1, Z_2) = \frac{1}{2}(X_1^{(1)} - X_2^{(1)})(X_1^{(2)} - X_2^{(2)})$. In this case, to construct the CAM estimator, we first use the simple choice $\phi_m(Z_1^m, Z_2^m) = (X_1^{(2)} - X_2^{(2)})^2$. Then, similarly to the previous problem we consider

$$\phi_m(Z_1^m, Z_2^m) = \{(\beta_1(Y_1 - Y_2) + \beta_2(X_1^{(2)} - X_2^{(2)}) + \beta_4(X_1^{(2)}Y_1 - X_2^{(2)}Y_2)\}(X_1^{(2)} - X_2^{(2)}), \tag{7}$$

where, again, the idea is to approximate the optimal $\phi_m^*(Z_1^m, Z_2^m) = \mathbb{E}\{\phi(Z_1, Z_2)|Z_1^m, Z_2^m\}$.

We compare the performance to the complete-case estimator with both versions of the CAM U-Statistic in Table 1. In each case, we present the point estimate, an approximate 95% confidence interval based on the result in Theorem 3 and the corresponding interval width. We see that the CAM estimator has a much narrower interval width in both problems. Moreover, in problem (i) the two CAM approaches lead to identical results (up to 3 significant figures), whereas in the second problem, the second CAM U-Statistic performs slightly better.

Method	Point est.	95% CI	CI width
(i) Marginal mean of bmi score			
Complete-case	0.55	(0.38, 0.73)	0.35
CAM: $\phi_m(Z^m) = X^{(2)}$	0.55	(0.44, 0.66)	**0.22**
CAM: $\phi_m(Z^m)$ linear – see (6)	0.55	(0.44, 0.66)	**0.22**
(ii) Covariance between height and weight			
Complete-case	1.27	(0.50, 2.04)	1.54
CAM: $\phi_m(Z_1^m, Z_2^m) = \frac{1}{2}(X_1^{(2)} - X_2^{(2)})^2$	1.17	(0.66, 1.68)	1.02
CAM: $\phi_m(Z_1^m, Z_2^m)$ linear – see (7)	1.19	(0.70, 1.69)	**0.99**

Table 1: Comparison of the complete-case and CAM U-Statistics using the Terneuzen birth cohort dataset

4. Nonparametric statistical learning

We now study two fundamental statistical learning problems, namely density estimation and regression. Typically in these problems, we are interested in estimating a function from $\mathbb{R}^d$ to $\mathbb{R}$, we show how the CAM estimator can be applied locally, i.e. for each $x \in \mathbb{R}^d$. Throughout this section we assume that the data is missing completely at random.

Density estimation and regression are canonical problems in statistics, and many nonparametric approaches have been proposed and studied in detail – see, for instance, Rosenblatt (1956), Parzen (1962), Wahba (1990), Wand and Jones (1995), Fan and Gijbels (1996), Carroll et al. (1998), Tsybakov (2004) and Biau and Devroye (2015). We focus our study on kernel based methods.

4.1 Kernel density estimation

In this subsection, assume we only observe $X_1^{m_1}, \ldots, X_n^{m_n}$ and we are interested in estimating f_X, the density of the marginal distribution of X. We specialise the setting introduced in Section 2 by letting $\theta(P) = f_X(x)$, the marginal density of X at a fixed $x \in \mathbb{R}^d$.

Let $h > 0$ be the bandwidth and, for $m \in \mathcal{M}$, let $K_m : \mathbb{R}^{d_m} \to [0, \infty)$ be a d_m-dimensional Kernel function. For $x \in \mathbb{R}^d$, $A \subseteq \{1, \ldots, n\}$ and $m \in \mathcal{M}$, let

$$\hat{f}_{A,m}(x^m) = \hat{f}_{A,m,h,K_m}(x^m) := \frac{1}{|A|h^{d_m}} \sum_{i \in A} K_m\Big(\frac{X_i^m - x^m}{h}\Big). \tag{8}$$

This can be thought of as an estimator of the marginal density $f_{X^m}(x^m)$ of X^m at x^m. In particular, the complete-case estimator of $f_X(x)$ is $\hat{f}_0 := \hat{f}_{A_0,0,h,K}(x)$, where $K = K_0$ is the d-dimensional kernel. Our CAM density estimator is constructed using $\hat{f}_0$, as well as $\hat{f}_{0,m} := \hat{f}_{A_0,m,h,K_m}(x^m)$ and $\hat{f}_m := \hat{f}_{A_m,m,h,K_m}(x^m)$, for $m \in \mathcal{M}$. To understand our choice of $\hat{f}_{0,m}$ and $\hat{f}_m$, recall from the discussion after Proposition 1 that the optimal choice of $\hat{f}_{0,m}$ is

$$\hat{f}^*_{0,m} = \mathbb{E}(\hat{f}_0 | \mathcal{T}_{A_0,m}) = \frac{1}{n_0 h^d} \sum_{i \in A_0} \mathbb{E}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big| X_i^m\Big\}. \tag{9}$$

This takes a similar form as the kernel density estimator in (8). At the end of this subsection we will see that $\hat{f}^*_{0,m}$ can often be well approximated using a local constant estimator with a practical choice of kernel K_m that depends on K. Finally, we have also chosen $\hat{f}_m$ so that $\mathbb{E}(\hat{f}_m) = \mathbb{E}(\hat{f}_{0,m})$.

Now let $\hat{f}_{0,\mathcal{M}} := (\hat{f}_{0,m} : m \in \mathcal{M})^T \in [0,\infty)^{|\mathcal{M}|}$, and $\hat{f}_{\mathcal{M}} := (\hat{f}_m : m \in \mathcal{M})^T \in [0,\infty)^{|\mathcal{M}|}$. Then, for $\gamma \in \mathbb{R}^{|\mathcal{M}|}$, define the CAM kernel density estimator

$$\hat{f}^{\mathcal{M}}_\gamma = \hat{f}^{\mathcal{M}}_\gamma(x) := \hat{f}_0 - \gamma^T(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}}).$$

Our theoretical results in this section will make use of conditions **A1** and **A2** given in Section A.4. The Lipshitz assumption on f_X and f_{X^m} in **A1** allows us to approximate the accuracy of the kernel density estimates of f_X and f_{X^m}, respectively. Whereas the assumption on the Kernel functions in **A2** is satisfied by many commonly used kernels.

For an estimate $\hat{f}$ of $f_X(x)$, let $\mathrm{MSE}(\hat{f}) = \mathrm{MSE}(\hat{f})(x) := \mathbb{E}[\{\hat{f} - f_X(x)\}^2]$. Under our assumptions, it is well-known that the complete-case estimator satisfies $\mathrm{MSE}(\hat{f}_0) = O(1/(n_0h^d) + h^2)$ as $n_0 \to \infty$; see, for example, Tsybakov (2004, Propositions 1.1 and 1.2) for the $d = 1$ case.

Let $\nu = \nu(K) := \int_{\mathbb{R}^d} K^2(z)\,dz < \infty$, and, for each $m \in \mathcal{M}$, let $\nu_m = \nu_m(K_m) := \int_{\mathbb{R}^{d_m}} K_m^2(z)\,dz < \infty$. Furthermore, for $m_1 \neq m_2 \in \mathcal{M}$, let $m^{1,2} = \mathrm{pmax}\{m_1, m_2\} \in \{0,1\}^d$ and $m_{1,2} = \mathrm{pmin}\{m_1, m_2\} \in \{0,1\}^d$ denote the entrywise maximums and minimums, respectively, of m_1 and m_2. For $m_1 \neq m_2$, let $\nu_{m_1,m_2} = \nu_{m_1,m_2}(K_{m_1}, K_{m_2}) := \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(z^{m_1})K_{m_2}(z^{m_2})\,dz^{m_{1,2}}$.

Our next result shows that the asymptotic difference between the mean squared error of our proposal and the complete-case estimator can be written in terms of γ, the $|\mathcal{M}|$-dimensional vector $\Omega_{\mathrm{D}} := (\frac{\nu_{0,m} f_X(x)}{n_0 h^{d_m}} : m \in \mathcal{M})^T$ and Λ_{D}, the symmetric $|\mathcal{M}| \times |\mathcal{M}|$ matrix with entries

$$\Lambda_{\mathrm{D},m,m} := \frac{\nu_m f_{X^m}(x^m)}{h^{d_m}}\Big(\frac{1}{n_0} + \frac{1}{n_m}\Big); \quad \Lambda_{\mathrm{D},m_1,m_2} := \frac{\nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}})}{n_0 h^{d_{m^{1,2}}}}, \tag{10}$$

for $m, m_1 \neq m_2 \in \mathcal{M}$.

Theorem 5 *Assume **A1** and **A2**. For $0 < \alpha < \beta < 1/d$, we have*

$$\mathrm{MSE}(\hat{f}^{\mathcal{M}}_\gamma) - \mathrm{MSE}(\hat{f}_0) = \big(\gamma^T\Lambda_{\mathrm{D}}\gamma - 2\gamma^T\Omega_{\mathrm{D}}\big)\{1 + o(1)\}$$

as $n \to \infty$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

Recall that by Proposition 1, the optimal γ, which maximises the improvement in mean squared error over the complete-case estimator, is

$$\gamma^* = \Lambda^{-1}\Omega = \Lambda_{\mathrm{D}}^{-1}\Omega_{\mathrm{D}}\{1 + O(h)\}. \tag{11}$$

Further, suppose that $\mathcal{M} = \{m\}$, then the corresponding maximum mean squared error reduction can be derived to be

$$\begin{aligned}\mathrm{MSE}(\hat{f}_0) - \mathrm{MSE}(\hat{f}^{\mathcal{M}}_{\gamma^*}) &= \frac{\mathrm{Cov}^2(\hat{f}_0, \hat{f}_{0,m})}{\mathrm{Var}(\hat{f}_{0,m}) + \mathrm{Var}(\hat{f}_m)} \\ &= \frac{n_m\nu_{0,m}^2 f_X(x) f_{X|X^m}(x|x^m)}{n_0 h^{d_m}(n_0 + n_m)\nu_m}\{1 + O(h)\} = O\Big(\frac{1}{n_0 h^{d_m}}\Big),\end{aligned}$$

as $n \to \infty$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Here $f_{X|X^m}(x|x^m)$ denotes the conditional density of X at x given $X^m = x^m$. We see that a larger improvement is possible when n_m is large compared to n_0, or if $f_{X|X^m}(x|x^m)$ is large. Note, however, that we only obtain a second order improvement over the complete-case approach. To understand this further, in contrast to the U-Statistics setting, in the density estimation problem $\hat{f}_0$ and $\hat{f}_{0,m} - \hat{f}_m$ have different convergence rates, with the later converging at a faster rate because of the smaller dimension d_m. Therefore the covariance between $\hat{f}_0$ and $\hat{f}_{0,m} - \hat{f}_m$ is negligible compared to the asymptotic variance of $\hat{f}_0$. Nevertheless, we will see in our numerical study in Section 4.3, that the improvement CAM offers over the complete-case method is appreciable in finite sample problems. Of course, Ω and Λ are unknown. Nonetheless, we have an immediate corollary that for any γ such that $\gamma^T \Lambda \gamma < 2\gamma^T \Omega$ the corresponding CAM estimator will lead to an improvement over the complete-case approach.

It remains to propose practical choices of tuning parameters. First, we suppose that the complete-case kernel K and bandwidth h are given to us; if needed these can be chosen using cross-validation on the complete cases. Now, to choose γ we attempt to approximate the optimal choice in (11) above. More precisely, let $\hat{\gamma}_{\mathrm{D},m} = \frac{\nu_{0,m} n_m \hat{f}_0}{\nu_m (n_0 \hat{f}_{0,m} + n_m \hat{f}_m)}$, where we have used $\hat{f}_0$ and $\frac{n_0 \hat{f}_{0,m} + n_m \hat{f}_m}{n_0 + n_m}$ as estimates of $f_X(x)$ and $f_{X^m}(x^m)$, respectively. We also approximate the off-diagonal terms in Λ_{D} by 0, since they are of smaller order than the terms on the diagonal, i.e. $\Lambda_{\mathrm{D}} = \mathrm{diag}(\Lambda_{\mathrm{D}})\{1 + o(1)\}$ – see (10). Finally, then, we let $\hat{\gamma}_{\mathrm{D}} := (\{\hat{\gamma}_{\mathrm{D},m} : m \in \mathcal{M}\})^T$.

We choose the kernel K_m in an attempt to mimic the optimal choice in (9). Lemma 8 in Section A.4 shows that for a large family of kernels, under appropriate smoothness conditions on $f_{X|X^m}$, we can approximate $\hat{f}^*_{0,m}$ up to first order using a kernel density estimator with practical choice of kernel K_m that depends only on K. For instance, if K is the Gaussian kernel $K(t) = \frac{1}{(2\pi)^{d/2}} \exp(-\|t\|^2/2)$, for $t \in \mathbb{R}^d$, then $\hat{f}^*_{0,m}$ is well-approximated by using the d_m-dimensional Gaussian kernel $K_m(z) = \frac{1}{(2\pi)^{d_m/2}} \exp(-\|z\|^2/2)$, for $z \in \mathbb{R}^{d_m}$, in (8).

4.2 Local constant regression

We now consider the standard homoscedastic nonparametric regression problem, where the pair (X, Y) takes values in $\mathbb{R}^d \times \mathbb{R}$ and satisfies the relationship

$$Y = \eta(X) + \sigma\epsilon.$$

Here $\sigma > 0$ and $\eta : \mathbb{R}^d \to \mathbb{R}$ is the regression function, i.e. $\eta(x) := \mathbb{E}(Y|X = x)$. The random variable ϵ has mean zero and variance one, and is independent of X. We are interested in estimating $\theta(P) = \eta(x)$, the regression function at a fixed $x \in \mathbb{R}^d$.

Consider also the regression model when X is missing the features $m \in \{0,1\}^d$. It is convenient to define $\eta_m(x^m) := \mathbb{E}(Y|X^m = x^m)$, and $\tau_m(x^m) := \mathrm{Var}\{\eta(X)|X^m = x^m\}$, where $\tau_0(\cdot) = \mathrm{Var}\{\eta(X)|X = x\} = 0$. Finally, for $m_1, m_2 \in \mathcal{M}$, let $\tau_{m_1,m_2}(x^{m_{1,2}}) := \mathbb{E}[\{\eta(X) - \eta_{m_1}(x^{m_1})\}\{\eta(X) - \eta_{m_2}(x^{m_2})\}|X^{m_{1,2}} = x^{m_{1,2}}]$.

Recall the bandwidth $h > 0$ and kernel function K_m, for $m \in \mathcal{M}$, used in the previous section. For $x \in \mathbb{R}^d$ and $A \subseteq \{1, \dots, n\}$, the *local constant estimator* of $\eta_m(x^m)$ is

$$\hat{\eta}_{A,m}(x^m) = \hat{\eta}_{A,m,h,K}(x^m) := \operatorname*{argmin}_{\alpha \in \mathbb{R}} \Big\{ \sum_{i \in A} K_m\Big(\frac{X_i^m - x^m}{h}\Big)(Y_i - \alpha)^2 \Big\}. \tag{12}$$

In particular, the complete-case estimator of $\eta(x)$ is $\hat{\eta}_0 := \hat{\eta}_{A_0,0,h,K}(x)$, where $K = K_0$ is a d-dimensional kernel. Further, let $\hat{\eta}_{0,m} := \hat{\eta}_{A_0,m,h,K_m}(x)$ and $\hat{\eta}_m := \hat{\eta}_{A_m,m,h,K_m}(x)$; in contrast to the density estimation setting, it is less clear why the form of $\hat{\eta}_{0,m}$ is effective here – we postpone discussion of this until the end of this subsection.

Let $\hat{\eta}_{0,\mathcal{M}} = (\hat{\eta}_{0,m} : m \in \mathcal{M})^T$ and $\hat{\eta}_{\mathcal{M}} = (\hat{\eta}_m : m \in \mathcal{M})^T$. Then, for $\gamma \in \mathbb{R}^{|\mathcal{M}|}$, we define the CAM local constant regression estimator

$$\hat{\eta}_\gamma^{\mathcal{M}} = \hat{\eta}_\gamma^{\mathcal{M}}(x) := \hat{\eta}_0 - \gamma^T(\hat{\eta}_{0,\mathcal{M}} - \hat{\eta}_{\mathcal{M}}).$$

Our main theoretical result in this section will make use of two further assumptions on the regression function; see **A3** and **A4** given in Section A.4. In particular, we ask that the functions η, η_m, τ_m and τ_{m_1,m_2} are Lipschitz, which means that these functions can be estimated efficiently using a local constant estimator. Now, for $m \in \mathcal{M}$, let $\mu_{0,m} = \mu_{0,m}(K_m) := \int_{\mathbb{R}^{d_m}} K_m(z)\, dz < \infty$. Further, let

$$\Omega_{\mathrm{R}} := \Big(\frac{\sigma^2 \nu_{0,m}}{\mu_{0,m} f_{X^m}(x^m) n_0 h^{d_m}} : m \in \mathcal{M} \Big)^T.$$

Let Λ_{R} be the $|\mathcal{M}| \times |\mathcal{M}|$ matrix with diagonal entries

$$\Lambda_{\mathrm{R},m,m} := \frac{\nu_m\{\sigma^2 + \tau_m(x^m)\}}{\mu_{0,m}^2 f_{X^m}(x^m) h^{d_m}} \Big(\frac{1}{n_0} + \frac{1}{n_m} \Big),$$

and off diagonal entries

$$\Lambda_{\mathrm{R},m_1,m_2} := \frac{\nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}}) \big\{ \sigma^2 + \tau_{m_1,m_2}(x^{m_{1,2}}) \big\}}{\mu_{0,m_1} f_{X^{m_1}}(x^{m_1}) \mu_{0,m_2} f_{X^{m_2}}(x^{m_2}) n_0 h^{d_{m^{1,2}}}}.$$

Note that the off-diagonal terms of Λ_{R} are of smaller order than the terms on the diagonal, i.e. $\Lambda_{\mathrm{R}} = \mathrm{diag}(\Lambda_{\mathrm{R}})\{1 + o(1)\}$. Finally, for a regression estimator $\hat{\eta}$ of $\eta(x)$, we write

$$\mathrm{MSE}(\hat{\eta}) = \mathrm{MSE}(\hat{\eta})(x) := \mathbb{E}[\{\hat{\eta} - \eta(x)\}^2 | Z_1^{m_1}, \dots, Z_n^{m_n}].$$

Theorem 6 *Assume **A1**, **A2**, **A3** and **A4**. Then, for each $0 < \alpha < \beta < 1/d$, we have*

$$\mathrm{MSE}(\hat{\eta}_\gamma^{\mathcal{M}}) - \mathrm{MSE}(\hat{\eta}_0) = \big(\gamma^T \Lambda_{\mathrm{R}} \gamma - 2\gamma^T \Omega_{\mathrm{R}} \big)\{1 + o_p(1)\}$$

as $n \to \infty$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

Theorem 6 gives the leading order asymptotic difference in mean squared error between the CAM estimator and the complete-case estimator. We see that the optimal leading order improvement in this case is $\Omega_{\mathrm{R}}^T \Lambda_{\mathrm{R}}^{-1} \Omega_{\mathrm{R}}$, which can achieved by taking $\gamma = \gamma_{\mathrm{R}}^* := \Lambda_{\mathrm{R}}^{-1} \Omega_{\mathrm{R}}$. Note also that, similarly to (11), the optimal γ^* given by Proposition 1 satisfies, $\gamma^* = \gamma_{\mathrm{R}}^* \{1 + O(h)\}$.

Again, in practice, we attempt to mimic the performance of the optimal estimator. First, ignoring the off-diagonal terms in Λ_{R}, we have

$$\gamma_{\mathrm{R}}^* \approx \Big(\frac{\sigma^2 \nu_{0,m} \mu_{0,m} n_m}{\{\sigma^2 + \tau_m(x^m)\} \nu_m (n_0 + n_m)} : m \in \mathcal{M} \Big)^T.$$

For the unknown terms, write $\sigma^2 = \mathbb{E}[\{Y - \eta(x)\}^2 | X = x]$ and $\sigma_m^2 := \sigma^2 + \tau_m(x^m) = \mathbb{E}[\{Y - \eta_m(x^m)\}^2 | X^m = x^m]$. These can be estimated in a natural way using the local constant method. In practice, the estimates $\hat{\sigma}^2$ and $\hat{\sigma}_m^2$ say can be calculated directly by reusing the weights from the regression estimators; cf. (22). Finally, this leads to a practical choice of

$$\hat{\gamma}_{\mathrm{R}} = \Big(\frac{\hat{\sigma}^2 \nu_{0,m} \mu_{0,m} n_m}{\hat{\sigma}_m^2 \nu_m (n_0 + n_m)} : m \in \mathcal{M} \Big)^T.$$

As noted above, the choice of kernel K_m is less straightforward than in the density estimation setting. Here we can write $\hat{\eta}_0 = \sum_{i \in A_0} Y_i K\big(\frac{X_i - x}{h}\big) / \{\sum_{i \in A_0} K\big(\frac{X_i - x}{h}\big)\}$, and the optimal choice of $\hat{\eta}_{0,m}$ is

$$\hat{\eta}_{0,m}^* = \mathbb{E}(\hat{\eta}_0 | \mathcal{T}_{A_0,m}) = \sum_{i \in A_0} Y_i \mathbb{E}\Big\{ \frac{K\big(\frac{X_i - x}{h}\big)}{\sum_{j \in A_0} K\big(\frac{X_j - x}{h}\big)} \Big| \mathcal{T}_{A_0,m} \Big\}.$$

Lemma 9 in Section A.6 shows that, under certain conditions, this optimal choice is well-approximated by our practical choice $\hat{\eta}_{0,m}$ with kernel K_m depending only on K. In particular, if K is the d-dimensional Gaussian kernel, then $\hat{\eta}_{0,m}$ can be constructed as in (12) using the d_m-dimensional Gaussian kernel. We see in our numerical study, that our CAM approach with this practical choice of $\hat{\eta}_{0,m}$ does often lead to an appreciable improvement over the complete-case estimator.

4.3 Numerical examples

We now demonstrate the CAM density and regression estimators with some numerical examples. Consider the following models

(a) Density model 1: $X \sim N_2(0, \Sigma)$, where $\Sigma = 0.3 I_2 + 0.7(1_2 1_2^T)$, where $1_2 := (1, 1)^T$.

(b) Density model 2: $X \sim U(B_1(0))$, where $B_1(0) \subseteq \mathbb{R}^2$ denotes the unit disk centred at 0.

(c) Density model 3: $X \sim \frac{1}{4} U([-2, -1] \times [-1/2, 1/2]) + \frac{3}{4} U([1, 2] \times [-1/2, 1/2])$.

(d) Regression model 1: Let $X \sim U([0, 1]^3)$, and $Y = X^{(1)} + X^{(2)} + 0.1\epsilon$, where $\epsilon \sim N(0, 1)$ is independent of X.

(e) Regression model 2: Let $X \sim U([0, 1]^3)$, and $Y = (X^{(1)} - X^{(2)})^2 + 0.1\epsilon$, where $\epsilon \sim N(0, 1)$ is independent of X.

(f) Regression model 3: Let $X \sim N_2(0, \Sigma)$, where $\Sigma = 0.3I_2 + 0.8(1_2 1_2^T)$ and $Y = \sin(2X^{(1)}) + 0.3\epsilon$, where $\epsilon \sim N(0,1)$ is independent of X.

In each case, we generate a training set of size $n \in \{200, 500\}$, and then introduce missingness by removing first component of X independently with probability $p_1 \in \{0.25, 0.5, 0.75\}$. Therefore, in settings (a), (b), (c), and (f), when $d = 2$, we have on average $n(1-p_1)$ complete cases in A_0 and an average of np_1 observations in A_m, for $m = (1,0)$, whereas $A_{(0,1)^T}$ and $A_{(1,1)^T}$ are empty; in other words $\mathcal{M} = \{(1,0)^T\}$. For settings (d) and (e), where $d = 3$, we again have an average of $n(1-p_1)$ complete cases in A_0, but now an average of np_1 observations in $A_{(1,0,0)^T}$, all other sets A_m are empty and $\mathcal{M} = \{(1,0,0)^T\}$.

The kernel density estimators are computed using the `ks` package available from CRAN. In particular, we use the `kde` function with a Gaussian kernel, and the diagonal bandwidth matrices were chosen using the `Hpi.diag` function. In the regression settings, we make use of the `regpro` package available from `CRAN`.

To measure the performance, recall that for two density functions f and g, say, the *Total Variation* distance between f and g is $\mathrm{TV}(f,g) := \frac{1}{2}\int_{\mathbb{R}^d} |f(x) - g(x)|\, dx$. We present boxplots of

$$\frac{\mathrm{TV}(\hat{f}_0, f_X) - \mathrm{TV}(\hat{f}^{\mathcal{M}}_{\hat{\gamma}_{\mathrm{D}}}, f_X)}{\mathrm{TV}(\hat{f}_0, f_X)}. \tag{13}$$

In the regression problems, we use the mean integrated squared error: for an estimate $\hat{\eta}$ of η that is $\mathrm{MISE}(\hat{\eta}) := \int_{\mathbb{R}^d} \{\hat{\eta}(x) - \eta(x)\}^2 \, dP_X(x)$. Then we present boxplots of

$$\frac{\mathrm{MISE}(\hat{\eta}_0, \eta) - \mathrm{MISE}(\hat{\eta}^{\mathcal{M}}_{\hat{\gamma}_{\mathrm{R}}}, \eta)}{\mathrm{MISE}(\hat{\eta}_0, \eta)}. \tag{14}$$

We see in Figure 3 that the CAM estimator outperforms the CC estimator in terms of total variation distance. As expected the CAM estimator leads to a greater reduction in error as the missingness probability p_1 increases. On the other hand, as n increases, the relative reduction in error is smaller. Finally, note that there are a small number of repeats of the experiment in each case where the difference in mean squared error is negative – this is likely due to the fact that we need to estimate the γ using the observed data, there is also a small Monte Carlo error.

Figure 4 shows that the CAM local constant regression estimator outperforms the CC estimator. Similarly to the density estimation problems, the CAM estimator leads to a greater reduction in error as the missingness probability p_1 increases. Here we see further that for larger sample sizes, the improvement relative to using the full data set appears to decrease slightly – this is in agreement with our Theorem 6, since we only expect a second order improvement in asymptotic mean squared error.

Finally, as discussed in the introduction, a popular approach to overcome issues with missing data is to impute the missing values and treat the result as a full dataset. We now compare the CAM approach to density estimation and regression with a number of state of the art imputation methods. These include *mean imputation*, where the missing entries for each feature are imputed with the sample mean (of the observed values) for that feature (Little and Rubin, 2002, Section 4.2.1); *Predictive mean matching (PMM) imputation*, here the missing values are replaced by a value sampled from the observed data, where the

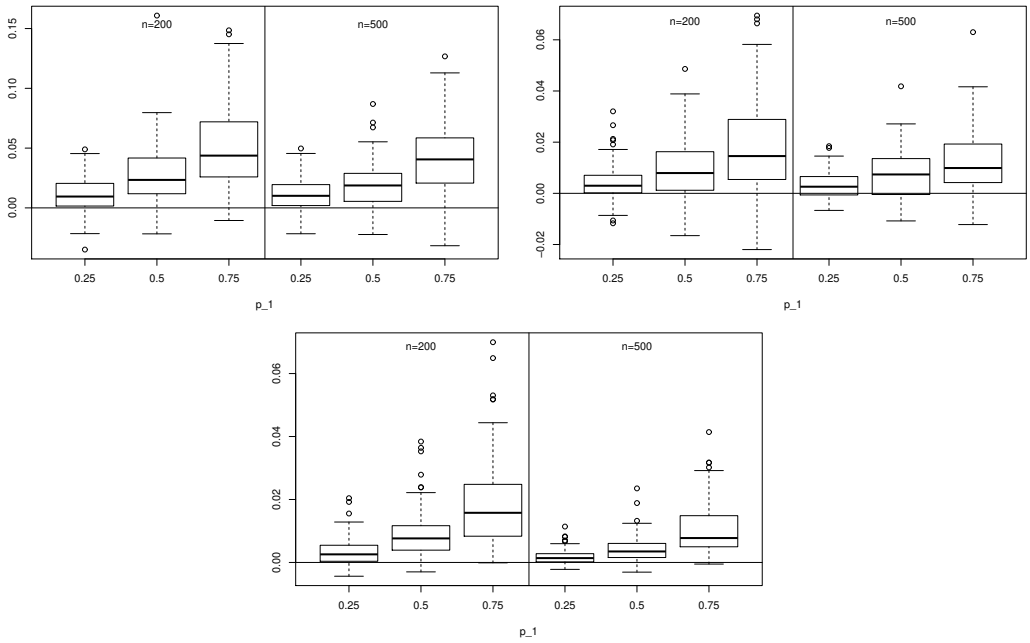

Figure 3: Boxplot of the relative performance of the CC and CAM kernel density estimators given by (13) for 100 repetitions of the experiment for density model 1 (top left), density model 2 (top right) and density model 3 (bottom).

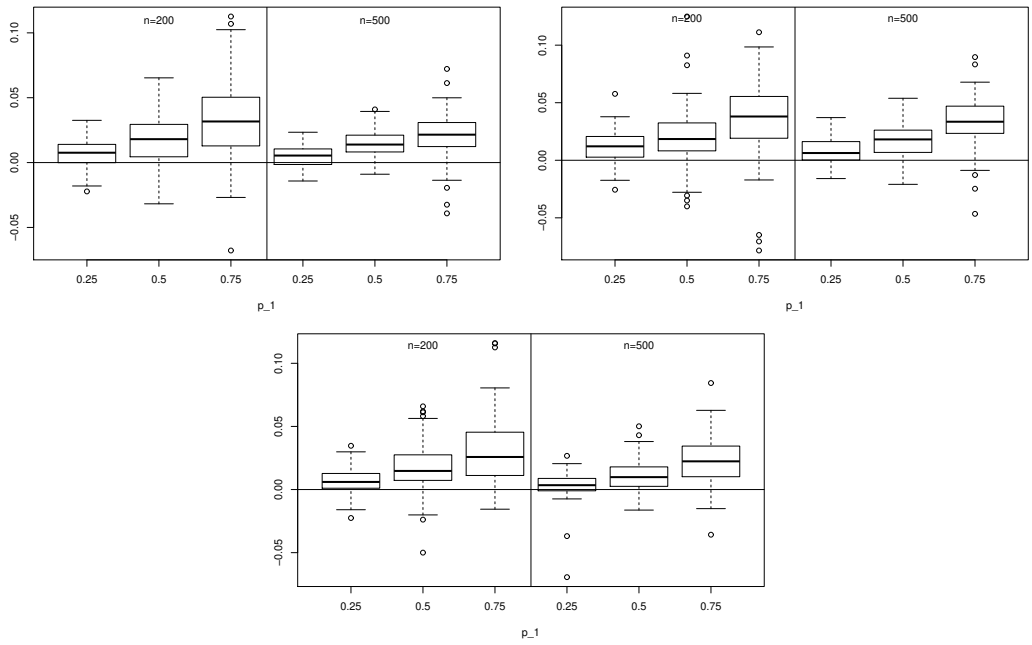

Figure 4: Boxplots of the relative performance of the CC and CAM local constant regression estimators given by (14) for 100 repetitions of the experiment for regression problem 1 (top left), regression problem 2 (top right) and regression problem 3 (bottom).

Model	Full	CAM	CC	Mean	PMM	RF
Density						
1	$4.97_{0.02}$	$6.07_{0.03}$	$6.23_{0.03}$	$23.77_{0.24}$	$6.76_{0.10}$	$6.86_{0.10}$
2	$35.31_{0.07}$	$40.28_{0.08}$	$40.58_{0.08}$	$101.80_{1.04}$	$50.21_{1.07}$	$38.31_{0.31}$
3	$51.84_{0.10}$	$52.49_{0.10}$	$52.72_{0.10}$	$69.66_{0.72}$	$52.91_{0.34}$	$52.85_{0.35}$
Regression						
1	$2.74_{0.01}$	$4.29_{0.02}$	$4.35_{0.02}$	$6.49_{0.10}$	$2.65_{0.04}$	$6.40_{0.11}$
2	$2.86_{0.02}$	$4.12_{0.03}$	$4.20_{0.03}$	$5.82_{0.11}$	$12.28_{0.23}$	$6.02_{0.12}$
3	$19.29_{0.11}$	$29.51_{0.19}$	$29.84_{0.20}$	$61.90_{1.01}$	$112.22_{1.63}$	$24.44_{0.46}$

Table 2: Average estimated total variation errors ($*10^3$) for the density estimation problems and mean integrated squared error ($*10^3$) for the regression problems. In each case, $n = 500$ and the first component of X is missing with probability 1/2.

sample is taken from the observations that are close to the observation with the missing value (Little and Rubin, 2002, Section 4.3.2); and *Random Forest (RF) imputation*, this uses the Random Forests algorithm to predict the missing entries based on the observed data (Breiman, 2002; Pantanowitz and Marwala, 2009). For the latter two methods, we use multiple imputation; the missing values are imputed five times and we then take the average of the results after fitting the model on the five imputed datasets. A detailed outline of these methods can be found in Little and Rubin (2002, Chapters 4 and 10). Our implementation utilises the `mice` R package available from CRAN (van Buuren et al., 2018).

In our experiments, the kernel density estimators are computed using the `ks` package available from CRAN. In particular, we use the `kde` function with a Gaussian kernel, and the diagonal bandwidth matrices were chosen using the `Hpi.diag` function. In the regression settings, we make use of the `regpro` package available from `CRAN`. For the imputation approaches, the corresponding kernel methods are applied to the imputed datasets.

In Table 2 we present the estimated total variation or mean integrated squared errors (with standard errors in subscript) over 100 repetitions of each experiment. For comparison, we also present the results of applying the kernel methods to the full dataset of size n. We see that the CAM technique improves on the complete-case approach in every setting. On the other hand, as discussed in the introduction, the imputation approaches are not always successful: indeed, there are many settings here where the imputation methods lead to worse performance than the complete-case method. Notice that occasionally the imputation methods, eg. random forests in Density Model 2 and predictive mean matching in Regression Model 1, perform very well (even outperforming the full data estimator in a few cases) – it is unclear why this is the case due to the black-box nature of these approaches.

4.4 The Brandsma school dataset

In this subsection, we show how the CAM local constant regression estimator can be used in practice with the `brandsma` dataset available in the MICE package. The full data set consists of 4106 observations of 14 features. We simplify the problem by retaining only 4 features, namely the "verbal IQ score", the "SES score", "language score pre", and "language score

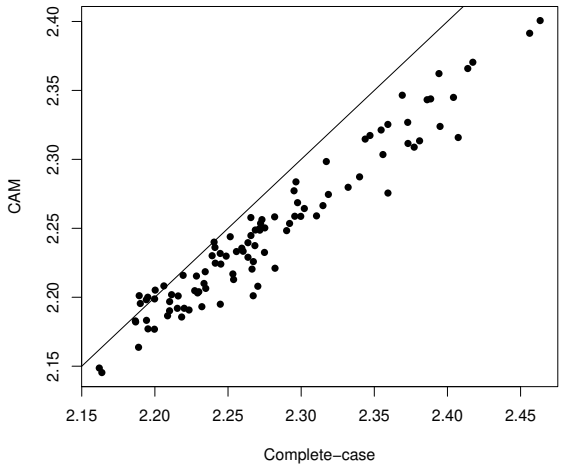

Figure 5: The average predictive MSE on the test set for the complete-case and CAM estimators for the Brandsma data application. The straight line is "$y = x$".

post". Suppose that we are interested in predicting the verbal IQ score, Y, from the remaining features, i.e. X is the 3-dimensional vector consisting of "SES score", "language score pre", and "language score post". We remove 17 observations for which the response is missing. In the resulting dataset, we have 3464 complete-cases, 302 with $m = m_1 := (0, 1, 0)^T$, 182 with $m = m_2 := (0, 0, 1)^T$, and 108 with $m = m_3 := (1, 0, 0)^T$. There are a few observations for other values of m, but the corresponding sample sizes are very small and are therefore ignored. Here we have $\mathcal{M} = \{m_1, m_2, m_3\}$.

In order to evaluate the performance of the CAM estimator, we take a subsample of size 1000 from the complete-cases to use as a test set (this is fixed throughout). We carry out 100 experiments. In each one, we form a training set by taking another sample of size 200 from the remaining 2464 complete-cases (this sample is different in each experiment). The 200 chosen complete-cases are then combined with the observations in A_{m_1}, A_{m_2} and A_{m_3} (which are the same in every experiment). Thus, in each experiment, we have $n_0 = 200$, $n_{m_1} = 302$, $n_{m_2} = 182$, and $n_{m_3} = 108$.

In Figure 5 we plot the average (over the test set) of $\{\hat{\eta}(X) - Y\}^2$ for the complete-case and CAM estimators in each of the 100 experiments. We use a Gaussian kernel and the bandwidth was chosen using leave-one-out cross-validation. We see that the CAM estimator has a lower predictive MSE than the complete-case in 94% of the cases. In the remaining 6% of cases, the performance of the two estimators is similar and the discrepancy can be explained by the small sample size used to estimate the MSE. The overall average MSE (with standard errors) of the complete-case and CAM estimators were 2.28 (0.007) and 2.25 (0.006), respectively, the average improvement in MSE is 0.03 (0.002), whereas the maximum improvement over the 100 experiments was 0.09.

5. Discussion

We have seen that our proposed CAM estimator can be used to improve the complete-case estimator in a wide range of statistical problems. We conclude our paper with a discussion of the computational cost and few extensions of our method.

First, the computational cost of our proposal consists of three main components. We need to compute complete-case estimator based on the data in A_0, as well as the correction term $\hat{\varphi}_{0,m} - \hat{\varphi}_m$, for each $m \in \mathcal{M}$. Computing the complete case estimator is an unavoidable step, and under our assumptions, cost of calculating $\hat{\varphi}_{0,m} - \hat{\varphi}_m$ will typically be of the same order as calculating the complete case estimator. Finally, we need to choose γ, which involves estimating the corresponding variance and covariance terms in γ^*. In many applications, when $|\mathcal{M}|$ is fixed, the main computational burden, then, may be in estimating γ^*. However, as we have demonstrated, there is often an efficient option available. In the U-statistics setting, the optimal γ is well approximated by using computationally feasible incomplete U-statistics as estimators; see the remarks immediately after the statement of Theorem 3. For the nonparametric density estimation problem, a data-driven choice of γ may be made at negligible extra computational cost by reusing our estimates of f_0, $f_{0,m}$ and f_m, and a similar technique may be used in nonparametric regression. When $|\mathcal{M}|$ is large, then the computational cost involved in using all $m \in \mathcal{M}$ may become burdensome and this may enter into the choice of $\mathcal{M}$.

A key aspect of our CAM proposal is to construct $\hat{\varphi}_{0,\mathcal{M}}$ and $\hat{\varphi}_{\mathcal{M}}$, so that $\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}$ is centered. One way to guarantee this is to use a resampling method. Let $n_{0,m} := \min\{n_0, n_m\}$. Consider the sets of all subsamples (without replacement) of $n_{0,m}$ observations from the data in A_0 and A_m, respectively, denoted by $A_0^1, \ldots, A_0^{B_0}$ and $A_m^1, \ldots, A_m^{B_m}$, where $B_0 = \binom{n_0}{n_{0,m}}$ and $B_m = \binom{n_m}{n_{0,m}}$. Notice that at least one of B_0 and B_m will be equal to one. Then, for $m \in \mathcal{M}$, consider the two (independent) statistics $\bar{\varphi}_{0,m} := \frac{1}{B_0}\sum_{b=1}^{B_0} \hat{\varphi}_{A_0^b,m}$ and $\bar{\varphi}_m := \frac{1}{B_m}\sum_{b=1}^{B_m} \hat{\varphi}_{A_m^b,m}$. At least in the missing completely at random setting, we have $\mathbb{E}(\bar{\varphi}_{0,m}) = \mathbb{E}(\bar{\varphi}_m)$ – each of the terms in the sums are estimators calculated on datasets of the same size. Let $\bar{\varphi}_{0,\mathcal{M}} = (\bar{\varphi}_{0,m} : m \in \mathcal{M})^T$, and $\bar{\varphi}_{\mathcal{M}} = (\bar{\varphi}_m : m \in \mathcal{M})^T$. Then, for $\gamma \in \mathbb{R}^{|\mathcal{M}|}$, define

$$\bar{\theta}_\gamma^{\mathcal{M}} := \hat{\theta}_0 - \gamma^T\big(\bar{\varphi}_{0,\mathcal{M}} - \bar{\varphi}_{\mathcal{M}}\big).$$

We have the following corollary to Proposition 1, which holds for any estimation method and requires no assumptions on the distribution P. The only restriction is that the data is MCAR.

Corollary 7 *Suppose the data is missing completely at random. Then*

$$\mathrm{MSE}(\bar{\theta}_\gamma^{\mathcal{M}}) - \mathrm{MSE}(\hat{\theta}_0) = \gamma^T \mathrm{Var}(\bar{\varphi}_{0,\mathcal{M}} - \bar{\varphi}_{\mathcal{M}})\gamma - 2\gamma^T \mathrm{Cov}(\hat{\theta}_0, \bar{\varphi}_{0,\mathcal{M}}). \tag{15}$$

Another consideration is the choice of the set $\mathcal{M}$. This choice is primarily driven by the data – in the first instance we might let $\mathcal{M} = \mathcal{M}^* := \{m \in \{0,1\}^d \setminus \{0_d\} : |A_m| > 0\}$. In some cases, however, we may consider using a different set $\mathcal{M}$. For instance, for some m, the corresponding sample size n_m may be non-zero but small. In our numerical analysis in Section 4.3, we drop m if $|A_m|/|A_0|$ is less than 0.1. Another potential option here is to use a data integration method. More specifically, let $\bar{A}_m := \{i \in A_0^c : m_i \leq m\}$, where the partial

order on $\{0,1\}^d$ is defined by $m_i \leq m$ if $\{j : m_i^j = 1\} \subseteq \{j : m^j = 1\}$. Then $\bar{A}_m$ is the largest set in A_0^c which has complete observations for variables in $\{j \in \{1, \dots d\} : m^j = 0\}$. Notice also that by construction $\bar{A}_m$ and A_0 are disjoint. To include the data integration in our CAM method, we can simply replace A_m with $\bar{A}_m$, for $m \in \mathcal{M}$. There are other options too. Suppose we have $m_1 \neq m_2 \in \{0,1\}^d \setminus \{0_d\}$, with $|A_{m_1}| > 0$, $|A_{m_2}| = 0$ and $m_1 \leq m_2$. Then rather than using A_{m_1}, we may opt to use $\bar{A}_{m_2} \supseteq A_{m_1}$ instead.

Finally we discuss the challenging non-MCAR settings. When the data is MAR or MNAR there is an additional challenge. In this case, the naive complete-case estimator will potentially be asymptotically biased. Conditionally on $M_1, \dots, M_n$, the data in $\mathcal{T}_{A_0,0}$ can be interpreted as independent and identically distributed pairs from Q_0, that is the joint distribution of a generic pair $(X,Y)|\{M = 0\}$. Thus, we expect the complete-case estimator $\hat{\theta}_0$ to be close to $\theta(Q_0)$ as opposed to $\theta(P_0)$. Two natural questions arise: (i) under what conditions do we have $\theta(P_0) = \theta(Q_0)$ and $\mathbb{E}(\hat{\varphi}_{0,m} - \hat{\varphi}_m) \approx 0$ (cf. Proposition 1)?; and (ii) if the conditions in (i) do not hold, how can we adapt the CAM estimator.

A partial answer to (i) is provided by the following in the regression setting: Suppose we observe n independent and identically distributed copies of $Z^M = (X^M, Y)$ and are interested in estimating $\eta(x) = \mathbb{E}(Y|X = x)$. Assume that M is independent of Y given X^M. (Note that this is slightly different to the missing at random condition, which assumes that M is independent of Z given Z^M.) In this case, we have that $\theta(Q_0) = \mathbb{E}(Y|X = x, M = 0) = \mathbb{E}(Y|X = x) = \theta(P_0)$ and

$$\theta_m(Q_m) = \mathbb{E}(Y|X^m = x^m, M = m) = \mathbb{E}(Y|X^m = x^m) = \mathbb{E}(Y|X^m = x^m, M = 0) = \theta_m(Q_0).$$

Thus, we can still expect that the complete-case approach will target $\eta(x)$, and moreover, can hope that $B_m = \mathbb{E}(\hat{\varphi}_{0,m} - \hat{\varphi}_m) \approx \theta_m(Q_0) - \theta_m(Q_m) = 0$.

Relating to the problem in (ii), in the missing at random case, there are many methods that aim to correct the bias of the naive complete-case estimator; see, for instance, Little and Rubin (2002, Chapter 3.3). One approach is to weight the observations according to the probability that they are (non)missing. For concreteness, we focus on one such idea, which advocates reweighting the observations according to their (inverse) propensity score – see Little and Rubin (2002, Chapter 3.7). Recall that $p_m(z) = p_m(x,y) = \mathbb{P}(M = m|X = x, Y = y)$. Of course, $p_m(x,y)$ is typically unknown, and needs to be estimated; in fact, we can only hope to estimate $p_m(x,y)$ from the observed data in the missing at random setting.

Consider the U-Statistics setting, where we are interested in estimating $\theta = \theta(P) = \mathbb{E}\{\phi(Z_1, \dots, Z_r)\}$. The complete-case approach we will in fact construct an unbiased estimator for $\theta(Q) = \mathbb{E}\big\{\phi(Z_1, \dots, Z_r)\big|M_1 = \dots = M_r = 0\big\} \neq \theta(P)$. One solution here is to consider the Horvitz-Thompson estimators (Horvitz and Thompson, 1956) in place of $\hat{\theta}_0$, $\hat{\varphi}_{0,m}$ and $\hat{\varphi}_m$. In this problem, the complete-case analogue is

$$\tilde{\theta}_0 = \tilde{\theta}_{A_0,0} = \frac{1}{\sum_{\{i_1,\dots,i_r\} \subseteq A_0} \frac{1}{\prod_{j=1}^r p_0(Z_{i_j})}} \sum_{\{i_1,\dots,i_r\} \subseteq A_0} \frac{\phi(Z_{i_1}, \dots, Z_{i_r})}{\prod_{j=1}^r p_0(Z_{i_j})}.$$

Moreover, similar expressions can be derived for $\tilde{\varphi}_{0,m}$ and $\tilde{\varphi}_m$. The CAM estimator can then be constructed as in the MCAR case. Of course, in practice, $p_0(z)$ and $p_m(z^m)$ need to be estimated using the observed data. This is non-trivial, and further study in this direction is left for future work.

Acknowledgements

We are grateful to the Editor and the two anonymous reviewers whose comments led to several improvements to the paper. The work of the first author was partially supported by an EPSRC New Investigator Award EP/V002694/1 and the work of the second author was partially supported by an NIH award 1R01GM131407.

Appendix A. Technical arguments

In this appendix, we present the proofs of all our theoretical results, as well as two auxiliary lemmas which justify our choice of the kernel estimators used in Section 4.

A.1 Proofs for Section 1.1

Proof [Proof of claims in Section 1.1] *Claim 1*: The statistic $T = (T_1, T_2)^T = (\hat{\nu}_{X,1} - \hat{\nu}_{Y,1}\Gamma_{12}/\Gamma_{22}, \hat{\nu}_{Y,1} + \hat{\nu}_{Y,2})^T$ is sufficient for $\nu = (\nu_X, \nu_Y)^T$. To see this we use the factorisation criteria: let $Z_i = (X_i, Y_i)^T$, the full likelihood is

$$\begin{aligned}
L(\nu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi|\Gamma|}} \exp\Big(-\frac{1}{2}(Z_i - \nu)^T\Gamma^{-1}(Z_i - \nu)\Big) \prod_{i=n+1}^{2n} \frac{1}{\sqrt{2\pi\Gamma_{22}}} \exp\Big(-\frac{1}{2\Gamma_{22}}(Y_i - \nu_Y)^2\Big) \\
&= \frac{1}{(2\pi)^n|\Gamma|^{n/2}\Gamma_{22}^{n/2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(Z_i - \nu)^T\Gamma^{-1}(Z_i - \nu) - \frac{1}{2\Gamma_{22}}\sum_{i=n+1}^{2n}(Y_i - \nu_Y)^2\Big) \\
&\propto \exp\Big(\sum_{i=1}^{n}\nu^T\Gamma^{-1}Z_i - \frac{1}{2}\nu^T\Gamma^{-1}\nu + \frac{1}{\Gamma_{22}}\sum_{i=n+1}^{2n} Y_i\nu_Y - \frac{nv_Y^2}{2\Gamma_{22}}\Big) \\
&= \exp\Big(n\nu^T\Gamma^{-1}(\hat{\nu}_{X,1}, \hat{\nu}_{Y,1})^T - \frac{1}{2}\nu^T\Gamma^{-1}\nu + \frac{1}{\Gamma_{22}}n\hat{\nu}_{Y,2}\nu_Y - \frac{nv_Y^2}{2\Gamma_{22}}\Big).
\end{aligned}$$

Now write

$$\nu^T\Gamma^{-1}(\hat{\nu}_{X,1}, \hat{\nu}_{Y,1})^T = \frac{1}{|\Gamma|}\Big\{(\Gamma_{22}\hat{\nu}_{X,1} - \Gamma_{12}\hat{\nu}_{Y,1})\nu_X + (\Gamma_{11}\hat{\nu}_{Y,1} - \Gamma_{12}\hat{\nu}_{X,1})\nu_Y\Big\}.$$

By collecting the coefficients of ν_X and ν_Y, it follows that

$$\Big(\hat{\nu}_{X,1} - \hat{\nu}_{Y,1}\frac{\Gamma_{12}}{\Gamma_{22}}, \hat{\nu}_{X,1} - \hat{\nu}_{Y,1}\frac{\Gamma_{11}}{\Gamma_{12}} - \hat{\nu}_{Y,2}\frac{\Gamma_{11}}{\Gamma_{12}} + \hat{\nu}_{Y,2}\frac{\Gamma_{12}}{\Gamma_{22}}\Big)^T = \Big(T_1, T_1 + (\frac{\Gamma_{12}}{\Gamma_{22}} - \frac{\Gamma_{11}}{\Gamma_{12}})T_2\Big)^T$$

is a sufficient statistic for ν. Thus Claim 1 is true.

Claim 2: We have that $\tilde{\nu}_X = \mathbb{E}(\hat{\nu}_{X,1}|T)$. To prove this, first observe that

$$(\hat{\nu}_{X,1}, \hat{\nu}_{Y,1}, \hat{\nu}_{Y,2})^T \sim N_3((\nu_X, \nu_Y, \nu_Y)^T, \Gamma'),$$

where

$$\Gamma' = \begin{pmatrix} \Gamma & 0 \\ 0 & \Gamma_{22} \end{pmatrix}.$$

Therefore

$$(\hat{\nu}_{X,1}, T_1, T_2)^T \sim N_3((\nu_X, \nu_X - \nu_Y\Gamma_{12}/\Gamma_{22}, 2\nu_Y)^T, \Gamma''),$$

where

$$\Gamma'' = \begin{pmatrix} \Gamma_{11} & \Gamma_{11} - \Gamma_{12}^2/\Gamma_{22} & \Gamma_{12} \\ \Gamma_{11} - \Gamma_{12}^2/\Gamma_{22} & \Gamma_{11} - \Gamma_{12}^2/\Gamma_{22} & 0 \\ \Gamma_{12} & 0 & 2\Gamma_{22} \end{pmatrix}.$$

By standard Gaussian distribution theory, it follows that

$$\begin{aligned} &\mathbb{E}\{\hat{\nu}_{X,1}|(T_1, T_2)\} \\ &= \nu_X + (\Gamma_{11} - \Gamma_{12}^2/\Gamma_{22}, \Gamma_{12}) \begin{pmatrix} \Gamma_{11} - \Gamma_{12}^2/\Gamma_{22} & 0 \\ 0 & 2\Gamma_{22} \end{pmatrix}^{-1} \begin{pmatrix} T_1 - \nu_X + \nu_Y\Gamma_{12}/\Gamma_{22} \\ T_2 - 2\nu_Y \end{pmatrix} \\ &= T_1 - T_2\Gamma_{12}/(2\Gamma_{22}) = \tilde{\nu}_X, \end{aligned}$$

which completes the proof. ■

A.2 Proofs of the results in Section 2

Proof of Proposition 1. First, we have that

$$\begin{aligned} \mathbb{E}\{(\hat{\theta}_\gamma^{\mathcal{M}} - \theta)^2 - (\hat{\theta}_0 - \theta)^2\} &= \mathbb{E}\{(\hat{\theta}_\gamma^{\mathcal{M}} - \hat{\theta}_0)^2 + 2\{\hat{\theta}_\gamma^{\mathcal{M}} - \hat{\theta}_0)(\hat{\theta}_0 - \theta)\} \\ &= \gamma^T\mathbb{E}\{(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})^T\}\gamma - 2\gamma^T\mathbb{E}\{(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})(\hat{\theta}_0 - \theta)\}. \end{aligned}$$

Then, using the fact that $\hat{\varphi}_{0,\mathcal{M}}$ and $\hat{\varphi}_{\mathcal{M}}$ are independent, write

$$\begin{aligned} &\mathbb{E}\{(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})^T\} \\ &\quad = \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}\hat{\varphi}_{0,\mathcal{M}}^T) - \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}\hat{\varphi}_{\mathcal{M}}^T) - \mathbb{E}(\hat{\varphi}_{\mathcal{M}}\hat{\varphi}_{0,\mathcal{M}}^T) + \mathbb{E}(\hat{\varphi}_{\mathcal{M}}\hat{\varphi}_{\mathcal{M}}^T) \\ &\quad = \mathrm{Var}(\hat{\varphi}_{0,\mathcal{M}}) + \mathrm{Var}(\hat{\varphi}_{\mathcal{M}}) + \{\mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}) - \mathbb{E}(\hat{\varphi}_{\mathcal{M}})\}\{\mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}) - \mathbb{E}(\hat{\varphi}_{\mathcal{M}})\}^T \\ &\quad = \mathrm{Var}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}}) + \{\mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}) - \mathbb{E}(\hat{\varphi}_{\mathcal{M}})\}\{\mathbb{E}(\hat{\varphi}_{0,\mathcal{M}}) - \mathbb{E}(\hat{\varphi}_{\mathcal{M}})\}^T. \end{aligned}$$

Moreover, since $\hat{\theta}_0$ and $\hat{\varphi}_{\mathcal{M}}$ are independent, we have

$$\begin{aligned} \mathbb{E}\{(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})(\hat{\theta}_0 - \theta)\} &= \mathbb{E}[(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})\{\hat{\theta}_0 - \mathbb{E}(\hat{\theta}_0) + \mathbb{E}(\hat{\theta}_0) - \theta\}] \\ &= \mathbb{E}[(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})\{\hat{\theta}_0 - \mathbb{E}(\hat{\theta}_0)\}] + \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})\mathbb{E}(\hat{\theta}_0 - \theta) \\ &= \mathrm{Cov}(\hat{\varphi}_{0,\mathcal{M}}, \hat{\theta}_0) + \mathbb{E}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})\mathbb{E}(\hat{\theta}_0 - \theta). \end{aligned}$$

The result follows. ■

A.3 Proofs of the results in Section 3

Proof of Theorem 2. Note that $\hat{\theta}_\gamma^{\mathcal{M}}$ can be written as

$$\hat{\theta}_\gamma^{\mathcal{M}} = (\hat{\theta}_0 - \gamma^T \hat{\varphi}_{0,\mathcal{M}}) + \gamma^T \hat{\varphi}_{\mathcal{M}}.$$

The two terms on the right hand side are independent because the former uses data in A_0 and the latter uses data in A_0^c. In addition, the first term on the right hand side is a U-Statistic with kernel

$$\tilde{\phi}(z_1, \cdots, z_r) = \phi(z_1, \cdots, z_r) - \sum_{m \in \mathcal{M}} \gamma_m \phi_m(z_1^m, \cdots, z_r^m),$$

and the second term is a linear combination of $\mathcal{M}$ independent U-Statistics each with kernel $\phi_m(z_1^m, \cdots, z_r^m)$. If $\mathbb{E}\{\phi^2(Z_1, \cdots, Z_r)\} < \infty$ and $\mathbb{E}\{\phi_m^2(Z_1^m, \cdots, Z_r^m)\} < \infty$ for all $m \in \mathcal{M}$, we have

$$\mathbb{E}\Big\{\phi(Z_1, \cdots, Z_r) - \sum_{m \in \mathcal{M}} \gamma_m \phi_m(Z_1^m, \cdots, Z_r^m)\Big\}^2 < \infty.$$

Let $\varphi_m = \mathbb{E}\{\phi_m\big(Z_1^m, \ldots, Z_r^m\big)\}$ and $\varphi_{\mathcal{M}} = (\varphi_m, m \in \mathcal{M})^T$. By classical U-Statistics theory (see for, example, van der Vaart (1998, Theorem 12.3)), we have that

$$\sqrt{n_0}\Big\{\hat{\theta}_0 - \gamma^T \hat{\varphi}_{0,\mathcal{M}} - (\theta - \gamma^T \varphi_{\mathcal{M}})\Big\} \to^d N(0, r^2 v_0); \quad \sqrt{n_m}(\hat{\varphi}_m - \varphi_m) \to^d N(0, r^2 v_m)$$

as $n \to \infty$, where $v_0 := \mathrm{Cov}\{\tilde{\phi}(Z_1, Z_2, \cdots, Z_r), \tilde{\phi}(Z_1, Z_{r+2}, \cdots, Z_{2r})\}$ and $v_m := \mathrm{Cov}\{\phi_m(Z_1^m, Z_2^m, \cdots, Z_r^m), \phi(Z_1^m, Z_{r+2}^m, \cdots, Z_{2r}^m)\}$.

Now, by the independence of $\hat{\theta}_0 - \gamma^T \hat{\varphi}_{0,\mathcal{M}}$ and $\hat{\varphi}_{\mathcal{M}}$, and noting that $\lim_{n\to\infty} \frac{n_m}{n_0} = \frac{q_m}{q_0}$, we have

$$\sqrt{n_0}\big(\hat{\theta}_\gamma^{\mathcal{M}} - \theta\big) \to^d N\Big(0, r^2\big(v_0 + \sum_{m \in \mathcal{M}} \gamma_m q_m^{-1} q_0 v_m\big)\Big).$$

Finally, by the definition of $\tilde{\phi}(z_1, \cdots, z_r)$ we can derive that $v_0 + \sum_{m \in \mathcal{M}} \gamma_m q_m^{-1} q_0 v_m = \psi_1 + \gamma^T \Lambda_1 \gamma - 2\gamma^T \Omega_1$. This completes the proof of the theorem. ∎

Proof of Theorem 3. First, we formally define our U-Statistic estimates of Ω_U and Λ_U. We have

$$\hat{\Omega}_{U,m} := \frac{1}{2\binom{n_0}{4r-2}} \sum_{\{i_1,\ldots,i_{4r-2}\} \subseteq A_0} \Big[\big\{\phi\big(Z_{i_1}, \ldots, Z_{i_r}\big) - \phi\big(Z_{i_{2r}}, \ldots, Z_{i_{3r-1}}\big)\big\} \big\{\phi_m\big(Z_{i_1}^m, Z_{i_{r+1}}^m, \ldots, Z_{i_{2r-1}}^m\big) - \phi_m\big(Z_{i_{2r}}^m, Z_{i_{3r}}^m, \ldots, Z_{i_{4r-2}}^m\big)\big\}\Big]. \tag{16}$$

Moreover

$$\hat{\Lambda}_{U,m,m} := \Big(1 + \frac{n_0}{n_m}\Big) \frac{1}{2\binom{n_0+n_m}{4r-2}} \sum_{\{i_1,\ldots,i_{4r-2}\} \subseteq A_0 \cup A_m} \Big[\big\{\phi_m\big(Z_{i_1}^m, \ldots, Z_{i_r}^m\big) - \phi_m\big(Z_{i_{2r}}^m, \ldots, Z_{i_{3r-1}}^m\big)\big\} \big\{\phi_m\big(Z_{i_1}^m, Z_{i_{r+1}}^m, \ldots, Z_{i_{2r-1}}^m\big) - \phi_m\big(Z_{i_{2r}}^m, Z_{i_{3r}}^m, \ldots, Z_{i_{4r-2}}^m\big)\big\}\Big]; \tag{17}$$

and

$$\hat{\Lambda}_{U,m_1,m_2} := \frac{1}{2\binom{n_0+n_{m_{1,2}}}{4r-2}} \sum_{\{i_1,\ldots,i_{4r-2}\}\subseteq A_0\cup A_{m_{1,2}}} \Big[\big\{\phi_{m_1}\big(Z_{i_1}^{m_1},\ldots,Z_{i_r}^{m_1}\big) - \phi_{m_1}\big(Z_{i_{2r}}^{m_1},\ldots,Z_{i_{3r-1}}^{m_1}\big)\big\} \big\{\phi_{m_2}\big(Z_{i_1}^{m_2},Z_{i_{r+1}}^{m_2},\ldots,Z_{i_{2r-1}}^{m_2}\big) - \phi_{m_2}\big(Z_{i_{2r}}^{m_2},Z_{i_{3r}}^{m_2},\ldots,Z_{i_{4r-2}}^{m_2}\big)\big\}\Big]. \quad (18)$$

Then, by classical U-Statistics theory (van der Vaart, 1998, Theorem 12.3), we can prove that $\hat{\Lambda}$ and $\hat{\Omega}$ are consistent estimators of Λ and Ω, respectively. Then the consistency of $\hat{\gamma}$ to γ^* follows automatically. Next note that

$$\sqrt{n_0}(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \theta) - \sqrt{n_0}(\hat{\theta}_{\gamma}^{\mathcal{M}} - \theta) = \sqrt{n_0}(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \hat{\theta}_{\gamma}^{\mathcal{M}}) = \sqrt{n_0}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})^T(\hat{\gamma} - \gamma^*).$$

By the proof of Theorem 2, the vector $\sqrt{n_0}(\hat{\varphi}_{0,\mathcal{M}} - \hat{\varphi}_{\mathcal{M}})$ is jointly asymptotically normal. Since $\hat{\gamma} - \gamma^* = o_p(1)$, it follows from the Slutsky's theorem that

$$\sqrt{n_0}(\hat{\theta}_{\hat{\gamma}}^{\mathcal{M}} - \theta) - \sqrt{n_0}(\hat{\theta}_{\gamma}^{\mathcal{M}} - \theta) \to^p 0.$$

This, together with Theorem 2, completes the proof of the theorem. ■

A.4 Conditions and proofs for the results in Section 4

We first formally state our assumptions. In the following, $L > 0$ is some universal constant.

A1 *Suppose that P_X and, for $m \in \{0,1\}^d \backslash \{(1,\ldots,1)^T\}$, the marginal X^m distribution P_{X^m} have densities f_X, and f_{X^m}, respectively, that satisfy $|f_X(z_1) - f_X(z_2)| \le L\|z_1 - z_2\|$, for all $z_1, z_2 \in \mathbb{R}^d$, and, for each $m \in \{0,1\}^d \setminus \{(1,\ldots,1)^T\}$, we have $|f_{X^m}(z_1^m) - f_{X^m}(z_2^m)| \le L\|z_1^m - z_2^m\|$, for all $z_1^m, z_2^m \in \mathbb{R}^{d_m}$.*

A2 *Suppose the kernel is such that $\bar{K} := \sup_{z\in\mathbb{R}^d}(1 + \|z\|)K(z) < \infty$, and that $\mu_0 = \mu_0(K) := \int_{\mathbb{R}^d} K(z)\,dz = 1$, $\mu_1 = \mu_1(K) := \int_{\mathbb{R}^d} \|z\|K(z)\,dz < \infty$. We also ask that $\nu = \nu(K) := \int_{\mathbb{R}^d} K^2(z)\,dz < \infty$. Moreover, for each $m \in \mathcal{M}$, we have $\bar{K}_m := \sup_{z\in\mathbb{R}^{d_m}}(1 + \|z\|)K_m(z) < \infty$, $\mu_{0,m} = \mu_{0,m}(K_m) := \int_{\mathbb{R}^{d_m}} K_m(z)\,dz < \infty$, $\mu_{1,m} = \mu_{1,m}(K_m) := \int_{\mathbb{R}^{d_m}} \|z\|K_m(z)\,dz < \infty$, and $\nu_m = \nu_m(K_m) := \int_{\mathbb{R}^{d_m}} K_m^2(z)\,dz < \infty$. Finally for $m_1 \ne m_2 \in \mathcal{M}$, letting $m^{1,2} = \text{pmax}\{m_1, m_2\} \in \{0,1\}^d$ and $m_{1,2} = \text{pmim}\{m_1, m_2\} \in \{0,1\}^d$ denote the entrywise maximums and minimums, respectively, of m_1 and m_2, finally we suppose that $\nu_{m_1,m_2} = \nu_{m_1,m_2}(K_{m_1}, K_{m_2}) := \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(z^{m_1})K_{m_2}(z^{m_2})\,dz^{m_{1,2}} < \infty$.*

A3 *We have that $|\eta(z_1) - \eta(z_2)| \le L\|z_1 - z_2\|$, for all $z_1, z_2 \in \mathbb{R}^d$, and, for each $m \in \mathcal{M}$, $|\eta_m(z_1^m) - \eta_m(z_2^m)| \le L\|z_1^m - z_2^m\|$, for all $z_1^m, z_2^m \in \mathbb{R}^{d_m}$.*

A4 *For each $m \in \mathcal{M}$, we have $|\tau_m(z_1^m) - \tau_m(z_2^m)| \le L\|z_1^m - z_2^m\|$, for all $z_1^m, z_2^m \in \mathbb{R}^{d_m}$. Finally, we ask, for all $m_1, m_2 \in \mathcal{M}$, that $|\tau_{m_1,m_2}(z_1^{m_{1,2}}) - \tau_{m_1,m_2}(z_2^{m_{1,2}})| \le L\|z_1^{m_{1,2}} - z_2^{m_{1,2}}\|$, for all $z_1^{m_{1,2}}, z_2^{m_{1,2}} \in \mathbb{R}^{d_{m_{1,2}}}$.*

Proof of Theorem 5. First, we have

$$\begin{aligned}\mathbb{E}[\{\hat{f}_\gamma^{\mathcal{M}} - f_X(x)\}^2 - \{\hat{f}_0 - f_X(x)\}^2] \\ &= \mathbb{E}[(\hat{f}_\gamma^{\mathcal{M}}(x) - \hat{f}_0)^2 + 2(\hat{f}_\gamma^{\mathcal{M}} - \hat{f}_0)\{\hat{f}_0 - f_X(x)\}] \\ &= \gamma^T \mathbb{E}\{(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})^T\}\gamma - 2\gamma^T \mathbb{E}[(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})\{\hat{f}_0 - f_X(x)\}].\end{aligned}$$

Now observe that, for each $m \in \mathcal{M}$,

$$\mathbb{E}(\hat{f}_{0,m} - \hat{f}_m) = \frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} \mathbb{E}\Big\{K_m\Big(\frac{X_i^m - x^m}{h}\Big)\Big\} - \frac{1}{n_m h^{d_m}} \sum_{i \in A_m} \mathbb{E}\Big\{K_m\Big(\frac{X_i^m - x^m}{h}\Big)\Big\} = 0.$$

Thus, using also that $\hat{f}_{0,\mathcal{M}}$ and $\hat{f}_{\mathcal{M}}$ are independent, we can write

$$\begin{aligned}\mathbb{E}\{(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})^T\} \\ &= \mathbb{E}[\{\hat{f}_{0,\mathcal{M}} - \mathbb{E}(\hat{f}_{0,\mathcal{M}}) + \hat{f}_{\mathcal{M}} - \mathbb{E}(\hat{f}_{\mathcal{M}})\}\{\hat{f}_{0,\mathcal{M}} - \mathbb{E}(\hat{f}_{0,\mathcal{M}}) + \hat{f}_{\mathcal{M}} - \mathbb{E}(\hat{f}_{\mathcal{M}})\}^T] \\ &= \mathrm{Cov}(\hat{f}_{0,\mathcal{M}}) + \mathrm{Cov}(\hat{f}_{\mathcal{M}}). \qquad (19)\end{aligned}$$

It remains to show that $\mathrm{Cov}(\hat{f}_{0,\mathcal{M}}) + \mathrm{Cov}(\hat{f}_{\mathcal{M}}) = \Lambda_D(1 + o(1))$ and $\mathbb{E}[(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})\{\hat{f}_0 - f_X(x)\}] = \Omega_D(1 + o(1))$.

For the diagonal terms in the covariances in (19), we have

$$\begin{aligned}\mathrm{Var}(\hat{f}_{0,m}) + \mathrm{Var}(\hat{f}_m) &= \frac{1}{h^{2d_m}}\Big(\frac{1}{n_0} + \frac{1}{n_m}\Big)\mathrm{Var}\Big\{K_m\Big(\frac{X^m - x^m}{h}\Big)\Big\} \\ &= \frac{1}{h^{2d_m}}\Big(\frac{1}{n_0} + \frac{1}{n_m}\Big)\Big[\int_{\mathbb{R}^{d_m}} K_m^2\Big(\frac{z^m - x^m}{h}\Big) f_{X^m}(z^m)\, dz^m \\ &\qquad - \Big\{\int_{\mathbb{R}^{d_m}} K_m\Big(\frac{z^m - x^m}{h}\Big) f_{X^m}(z^m)\, dz^m\Big\}^2\Big].\end{aligned}$$

After making the substitution $u^m = \frac{z^m - x^m}{h}$ and using assumptions **A1** and **A2**, we deduce that

$$\begin{aligned}\Big|\int_{\mathbb{R}^{d_m}} K_m^2\Big(\frac{z^m - x^m}{h}\Big) f_{X^m}(z^m)\, dz^m - h^{d_m} f_{X^m}(x^m)\nu_m\Big| \\ &\le h^{d_m} \int_{\mathbb{R}^{d_m}} K_m^2(u^m)|f_{X^m}(x^m + hu^m) - f_{X^m}(x^m)|\, du^m \le Lh^{d_m+1}\bar{K}_m\mu_{1,m}.\end{aligned}$$

Whereas

$$\begin{aligned}\Big|\int_{\mathbb{R}^{d_m}} K_m\Big(\frac{z^m - x^m}{h}\Big) f_{X^m}(z^m)\, dz^m - h^{d_m} f_{X^m}(x^m)\mu_{0,m}\Big| \\ &\le h^{d_m} \int_{\mathbb{R}^{d_m}} K_m(u^m)|f_{X^m}(x^m + hu^m) - f_{X^m}(x^m)|\, du^m \le Lh^{d_m+1}\mu_{1,m}. \qquad (20)\end{aligned}$$

It follows that

$$\mathrm{Var}(\hat{f}_{0,m}) + \mathrm{Var}(\hat{f}_m) = \frac{f_{X^m}(x^m)\nu_m}{h^{d_m}}\Big(\frac{1}{n_0} + \frac{1}{n_m}\Big)\{1 + o(1)\},$$

as $n \to \infty$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

For the off-diagonal terms in (19): first, for $m_1 \neq m_2 \in \mathcal{M}$, we have that $\mathrm{Cov}(\hat{f}_{m_1}, \hat{f}_{m_2}) = 0$, since A_{m_1} and A_{m_2} are disjoint for $m_1 \neq m_2$. For the remaining terms, we have

$$\begin{aligned}
&\mathrm{Cov}(\hat{f}_{0,m_1}, \hat{f}_{0,m_2}) \\
&\quad = \frac{1}{n_0 h^{d_{m_1}+d_{m_2}}} \mathrm{Cov}\Big\{ K_{m_1}\Big(\frac{X^{m_1} - x^{m_1}}{h}\Big), K_{m_2}\Big(\frac{X^{m_2} - x^{m_2}}{h}\Big)\Big\} \\
&\quad = \frac{1}{n_0 h^{d_{m_1}+d_{m_2}}} \bigg\{ \int_{\mathbb{R}^d} K_{m_1}\Big(\frac{z^{m_1} - x^{m_1}}{h}\Big) K_{m_2}\Big(\frac{z^{m_2} - x^{m_2}}{h}\Big) f_X(z)\, dz \bigg\} \\
&\qquad - \frac{1}{n_0 h^{d_{m_1}+d_{m_2}}} \bigg\{ \int_{\mathbb{R}^d} K_{m_1}\Big(\frac{z^{m_1} - x^{m_1}}{h}\Big) f_X(z)\, dz \bigg\} \bigg\{ \int_{\mathbb{R}^d} K_{m_2}\Big(\frac{z^{m_2} - x^{m_2}}{h}\Big) f_X(z)\, dz \bigg\}.
\end{aligned}$$

As above, we make the substitution $u = \frac{z-x}{h}$, which gives

$$\begin{aligned}
&\int_{\mathbb{R}^d} K_{m_1}\Big(\frac{z^{m_1} - x^{m_1}}{h}\Big) K_{m_2}\Big(\frac{z^{m_2} - x^{m_2}}{h}\Big) f_X(z)\, dz \\
&\qquad = h^d \int_{\mathbb{R}^d} K_{m_1}(u^{m_1}) K_{m_2}(u^{m_2}) f_X(x + hu)\, du \\
&\qquad = h^d \int_{\mathbb{R}^{d_{m_{1,2}}}} \int_{\mathbb{R}^{d-d_{m_{1,2}}}} K_{m_1}(u^{m_1}) K_{m_1}(u^{m_2}) f_X(x + hu)\, du^{1_d - m_{1,2}} du^{m_{1,2}} \\
&\qquad = h^d \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1}) K_{m_2}(u^{m_2}) \int_{\mathbb{R}^{d-d_{m_{1,2}}}} f_X(x + hu)\, du^{1_d - m_{1,2}} du^{m_{1,2}} \\
&\qquad = h^{d_{m_{1,2}}} \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1}) K_{m_2}(u^{m_2}) f_{X^{m_{1,2}}}(x^{m_{1,2}} + hu^{m_{1,2}})\, du^{m_{1,2}} \\
&\qquad = h^{d_{m_{1,2}}} \nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}}) \{1 + o(1)\},
\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. It follows from the previous calculation and (20), that

$$\mathrm{Cov}(\hat{f}_{0,m_1}, \hat{f}_{0,m_2}) = \frac{\nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}})}{n_0 h^{d_{m^{1,2}}}} \{1 + o(1)\},$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. This proves that first claim that $\mathrm{Cov}(\hat{f}_{0,\mathcal{M}}) + \mathrm{Cov}(\hat{f}_{\mathcal{M}}) = \Lambda_D(1 + o(1))$.

Finally, since $\hat{f}_0$ and $\hat{f}_{\mathcal{M}}$ are independent, we have that

$$\begin{aligned}
\mathbb{E}[(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})\{\hat{f}_0 - f_X(x)\}] &= \mathbb{E}[(\hat{f}_{0,\mathcal{M}} - \hat{f}_{\mathcal{M}})\{\hat{f}_0 - \mathbb{E}(\hat{f}_0) + \mathbb{E}(\hat{f}_0) - f_X(x)\}] \\
&= \mathbb{E}[\{\hat{f}_{0,\mathcal{M}} - \mathbb{E}(\hat{f}_{0,\mathcal{M}}) - \hat{f}_{\mathcal{M}} + \mathbb{E}(\hat{f}_{\mathcal{M}})\}\{\hat{f}_0 - \mathbb{E}(\hat{f}_0)\}] \\
&= \mathrm{Cov}(\hat{f}_{0,\mathcal{M}}, \hat{f}_0) + \mathrm{Cov}(\hat{f}_{\mathcal{M}}, \hat{f}_0) = \mathrm{Cov}(\hat{f}_{0,\mathcal{M}}, \hat{f}_0).
\end{aligned}$$

Then, reusing the covariance calculation above, for $m \in \mathcal{M}$, we have

$$\mathrm{Cov}(\hat{f}_0, \hat{f}_{0,m}) = \frac{1}{n_0 h^{d+d_m}} \mathrm{Cov}\Big\{ K\Big(\frac{X - x}{h}\Big), K_m\Big(\frac{X^m - x^m}{h}\Big)\Big\} = \frac{\nu_{0,m} f_X(x)}{n_0 h^{d_m}} \{1 + o(1)\},$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. This completes the proof of the second claim and hence concludes the proof of the theorem. ∎

Proof of Theorem 6. First, conditionally on the observed data, we have

$$\begin{aligned}\mathbb{E}[\{\hat{\eta}_\gamma^{\mathcal{M}} - \eta(x)\}^2 | X_1^{m_1}, \ldots, X_n^{m_n}] - \mathbb{E}[\{\hat{\eta}_0 - \eta(x)\}^2 | X_1^{m_1}, \ldots, X_n^{m_n}] \\ = \mathbb{E}\{(\hat{\eta}_\gamma^{\mathcal{M}} - \hat{\eta}_0)^2 | X_1^{m_1}, \ldots, X_n^{m_n}\} + 2\mathbb{E}[\{\hat{\eta}_0 - \eta(x)\}\{\hat{\eta}_\gamma^{\mathcal{M}} - \hat{\eta}_0\} | X_1^{m_1}, \ldots, X_n^{m_n}]. \quad (21)\end{aligned}$$

We analyse the two terms in (21) separately. We first introduce some notation and facts that will be used repeatedly in the proof. Recall that we can write the local constant estimator as a linear function of the responses $Y_{[n]} := (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. Indeed, let

$$W_{A,m} = W_{A,m,h,K_m} := \left(K_m\Big(\frac{X_1^m - x^m}{h}\Big) \mathbb{1}_{\{1 \in A\}}, \ldots, K_m\Big(\frac{X_n^m - x^m}{h}\Big) \mathbb{1}_{\{n \in A\}} \right)^T. \quad (22)$$

Then $\hat{\eta}_{A,m}(x^m) = H_{A,m,h,K_m}^T Y_{[n]}$, where $H_{A,m} = H_{A,m,h,K_m} := (\sum_{i=1}^n W_{A,m,i})^{-1} W_{A,m}$. For simplicity of presentation, write $H_0 = H_{A_0,0}$, $H_{0,m} = H_{A_0,m}$ and $H_m = H_{A_m,m}$, and let $H_{0,\mathcal{M}} = (H_{0,m} : m \in \mathcal{M})$ be the $n \times |\mathcal{M}|$ matrix with columns $H_{0,m}$, for $m \in \mathcal{M}$, and similarly let $H_{\mathcal{M}} = (H_m : m \in \mathcal{M}) \in \mathbb{R}^{n \times |\mathcal{M}|}$.

Next, let $E = \mathbb{E}(Y_{[n]} | X_1^{m_1}, \ldots, X_n^{m_n}) = \big(\eta_{m_1}(X_1^{m_1}), \ldots, \eta_{m_n}(X_n^{m_n})\big)^T$ be the conditional expectation of the responses, and let $\Gamma := \mathrm{diag}\{\tau_{m_1}(X_1^{m_1}), \ldots, \tau_{m_n}(X_n^{m_n})\}$. Further, let $\Gamma^* = \mathrm{diag}\{\tau_{m_1}(x^{m_1}), \ldots, \tau_{m_n}(x^{m_n})\}$ and $E^* = \{\eta_{m_1}(x^{m_1}), \ldots, \eta_{m_n}(x^{m_n})\}^T$. We will make use of the following facts (i) $1_n^T H_{A,m} = 1$; (ii) $E^{*T} H_0 = \eta(x)$, (iii) $H_m^T E^* = \eta_m(x^m)$, (iv) $H_{0,m}^T E^* = \eta(x)$, (v) $H_m^T \Gamma^* H_m = \tau_m(x^m) H_m^T H_m$, (vi) $H_{0,m}^T \Gamma^* = 0$, (vii) $\Gamma H_0 = \Gamma^* H_0 = 0$ and (viii) $H_{\mathcal{M}}^T H_0 = 0$.

Furthermore, we claim that, for $m \in \mathcal{M}$, the following results are true

$$H_0^T H_{0,m} = \frac{\nu_{0,m}}{\mu_{0,m} f_{X^m}(x^m) n_0 h^{d_m}} \{1 + O_p(h)\}; \quad (23)$$

$$H_{0,m}^T H_{0,m} = \frac{\nu_m}{\mu_{0,m}^2 f_{X^m}(x^m) n_0 h^{d_m}} \{1 + O_p(h)\}; \quad (24)$$

and

$$H_m^T H_m = \frac{\nu_m}{\mu_{0,m}^2 f_{X^m}(x^m) n_m h^{d_m}} \{1 + O_p(h)\}, \quad (25)$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Furthermore, for $m_1 \neq m_2 \in \mathcal{M}$, we have

$$H_{0,m_1}^T H_{0,m_2} = \frac{\nu_{m_1,m_2} f_{X^{m_1,2}}(x^{m_1,2})}{\mu_{0,m_1} f_{X^{m_1}}(x^{m_1}) \mu_{0,m_2} f_{X^{m_2}}(x^{m_2}) n_0 h^{d_{m^1,2}}} \{1 + O_p(h)\}; \quad (26)$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. To see (23), write

$$\begin{aligned} H_0^T H_{0,m} &= \frac{\sum_{i \in A_0} K\big(\frac{X_i - x}{h}\big) K_m\big(\frac{X_i^m - x^m}{h}\big)}{\sum_{i \in A_0} K\big(\frac{X_i - x}{h}\big) \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)} \\ &= \frac{\frac{1}{n_0^2 h^{d+d_m}} \sum_{i \in A_0} K\big(\frac{X_i - x}{h}\big) K_m\big(\frac{X_i^m - x^m}{h}\big)}{\frac{1}{n_0 h^d} \sum_{i \in A_0} K\big(\frac{X_i - x}{h}\big) \frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)}. \end{aligned} \quad (27)$$

We consider the numerator and denominator above separately. For the numerator, observe that

$$\begin{aligned}
&\mathbb{E}\Big\{\frac{1}{n_0^2 h^{d+d_m}}\sum_{i\in A_0} K\big(\frac{X_i - x}{h}\big)K_m\big(\frac{X_i^m - x^m}{h}\big)\Big\} \\
&\qquad = \frac{1}{n_0 h^{d+d_m}}\int_{\mathbb{R}^d} K\big(\frac{z-x}{h}\big)K_m\big(\frac{z^m - x^m}{h}\big) f_X(z)\,dz \\
&\qquad = \frac{1}{n_0 h^{d_m}}\int_{\mathbb{R}^d} K(u)K_m(u^m) f_X(x+hu)\,dz \\
&\qquad = \frac{\nu_{0,m} f_X(x)}{n_0 h^{d_m}} + \frac{1}{n_0 h^{d_m}}\int_{\mathbb{R}^d} K(u)K_m(u^m)\{f_X(x+hu) - f_X(x)\}\,dz.
\end{aligned}$$

Moreover, by Assumptions **A1** and **A2**, we have

$$\int_{\mathbb{R}^d}\Big|K(u)K_m(u^m)\{f_X(x+hu) - f_X(x)\}\Big|\,dz \le Lh\int_{\mathbb{R}^d} K(u)K_m(u^m)\|u\|\,du \le Lh\mu_1\bar{K}_m.$$

Hence, using Markov's inequality, the numerator in (27) admits the following expression

$$\frac{1}{n_0^2 h^{d+d_m}}\sum_{i\in A_0} K\Big(\frac{X_i - x}{h}\Big)K_m\Big(\frac{X_i^m - x^m}{h}\Big) = \frac{\nu_{0,m} f_X(x)}{n_0 h^{d_m}} + O_p(1/(n_0 h^{d_m - 1})),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. For the denominator, by appealing to similar arguments to those in the proof of Theorem 5, we have that

$$\frac{1}{n_0 h^d}\sum_{i\in A_0} K\Big(\frac{X_i - x}{h}\Big) = f_X(x) + O_p(h),$$

and

$$\frac{1}{n_0 h^{d_m}}\sum_{i\in A_0} K_m\Big(\frac{X_i^m - x^m}{h}\Big) = f_{X^m}(x^m)\mu_{0,m} + O_p(h),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. The claim in (23) then follows by Slutsky's Theorem. Moreover, the claims in (24) and (25) follow by the same argument with only minor changes. To see (26), observe that

$$\begin{aligned}
&\mathbb{E}\Big\{\frac{1}{n_0^2 h^{d_{m_1}+d_{m_2}}}\sum_{i\in A_0} K_{m_1}\big(\frac{X_i^{m_1} - x^{m_1}}{h}\big)K_{m_2}\big(\frac{X_i^{m_2} - x^{m_2}}{h}\big)\Big\} \\
&\qquad = \frac{1}{n_0 h^{d_{m_1}+d_{m_2}}}\int_{\mathbb{R}^d} K_{m_1}\big(\frac{z^{m_1} - x^{m_1}}{h}\big)K_{m_2}\big(\frac{z^{m_2} - x^{m_2}}{h}\big) f_X(z)\,dz \\
&\qquad = \frac{1}{n_0 h^{d_{m_1}+d_{m_2}}}\int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}\big(\frac{z^{m_1} - x^{m_1}}{h}\big)K_{m_2}\big(\frac{z^{m_2} - x^{m_2}}{h}\big) f_{X^{m_{1,2}}}(z^{m_{1,2}})\,dz^{m_{1,2}} \\
&\qquad = \frac{1}{n_0 h^{d_{m_1}+d_{m_2}-d_{m_{1,2}}}}\int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1})K_{m_2}(u^{m_2}) f_{X^{m_{1,2}}}(x^{m_{1,2}} + hu^{m_{1,2}})\,du^{m_{1,2}} \\
&\qquad = \frac{\nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}})}{n_0 h^{d_{m^{1,2}}}} + O(1/(n_0 h^{d_{m^{1,2}}-1})),
\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. In the last step above we have used the fact that, by Assumptions **A1** and **A2**, we have

$$\begin{aligned}
&\int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1})K_{m_2}(u^{m_2})\{f_{X^{m_{1,2}}}(x^{m_{1,2}} + hu^{m_{1,2}}) - f_{X^{m_{1,2}}}(x^{m_{1,2}})\}\, du^{m_{1,2}} \\
&\quad \le Lh \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1})K_{m_2}(u^{m_2})\|u^{m_{1,2}}\|\, du^{m_{1,2}} \\
&\quad \le Lh \int_{\mathbb{R}^{d_{m_{1,2}}}} K_{m_1}(u^{m_1})K_{m_2}(u^{m_2})(\|u^{m_1}\| + \|u^{m_2}\|)\, du^{m_{1,2}} \le Lh(\mu_{1,m_1}\bar{K}_{m_2} + \mu_{1,m_2}\bar{K}_{m_1}).
\end{aligned}$$

The claim in (26) then follows by similar arguments used to prove (23). We now return to the main argument.

Part I: the first term in (21): By definition of $Y_{[n]}$ we have

$$\mathbb{E}(Y_{[n]}Y_{[n]}^T | X_1^{m_1}, \ldots, X_n^{m_n}) = EE^T + \sigma^2 I_{n\times n} + \Gamma.$$

It follows from the definitions of $\hat{\eta}_\gamma^{\mathcal{M}}$ and $\hat{\eta}_0$ that the first term in (21) takes the form

$$\begin{aligned}
\mathbb{E}\{(\hat{\eta}_\gamma^{\mathcal{M}} - \hat{\eta}_0)^2 | X_1^{m_1}, \ldots, X_n^{m_n}\} &= \gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big)(EE^T + \sigma^2 I_{n\times n} + \Gamma)\big(H_{0,\mathcal{M}} - H_{\mathcal{M}}\big)\gamma \\
&= \gamma^T \Lambda_R \gamma + R_1 + R_2, \qquad (28)
\end{aligned}$$

where

$$R_1 := \gamma^T H_{\mathcal{M}}^T (\Gamma - \Gamma^*) H_{\mathcal{M}} \gamma$$

and

$$R_2 := \{\gamma^T (H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T)E\}^2 + \sigma^2 \gamma^T H_{0,\mathcal{M}}^T H_{0,\mathcal{M}}\gamma + \gamma^T H_{\mathcal{M}}^T(\sigma^2 I + \Gamma^*)H_{\mathcal{M}}\gamma - \gamma^T \Lambda_{\mathrm{R}}\gamma.$$

Here we have used that $H_{0,\mathcal{M}}^T(\Gamma - \Gamma^*)H_{0,\mathcal{M}} = 0$. We next show that both R_1 and R_2 are small order terms.

To bound R_1: By assumption **A4**, we have that

$$|\tau_m(x_i^m) - \tau_m(x^m)| \le L\|x_i^m - x^m\|.$$

Moreover, similarly to (25), by Assumption **A2**, we have

$$\begin{aligned}
H_m^T \mathrm{diag}(\|X_i^{m_i} - x^{m_i}\|, i = 1, \ldots, n) H_m &= \frac{\sum_{i\in A_m} \|X_i^m - x^m\| K_m^2\big(\frac{X_i^m - x^m}{h}\big)}{\sum_{i\in A_m} K_m\big(\frac{X_i^m - x^m}{h}\big) \sum_{i\in A_m} K_m\big(\frac{X_i^m - x^m}{h}\big)} \\
&\le \frac{\frac{1}{n_m^2 h^{2d_m - 1}} \sum_{i\in A_m} \bar{K}_m K_m\big(\frac{X_i^m - x^m}{h}\big)}{\{\frac{1}{n_m h^{d_m}} \sum_{i\in A_m} K_m\big(\frac{X_i^m - x^m}{h}\big)\}^2} \\
&= O_p\big(\frac{1}{n_m h^{d_m - 1}}\big).
\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Thus

$$\begin{aligned}
|R_1| &= \Big|\gamma^T H_{\mathcal{M}}^T (\Gamma - \Gamma^*) H_{\mathcal{M}} \gamma\Big| \\
&\le \sum_{m\in\mathcal{M}} \gamma_m^2 H_m^T \mathrm{diag}(\|X_i^{m_i} - x^{m_i}\|, i = 1, \ldots, n) H_m = O_p\Big(\sum_{m\in\mathcal{M}} \frac{\gamma_m^2}{n_m h^{d_m - 1}}\Big),
\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

To bound R_2: First we decompose R_2 into the sum of three terms. To that end, let

$$R_{21} := \{\gamma^T(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T)E\}^2 - \sum_{m\in\mathcal{M}} \frac{\gamma_m^2\nu_m\tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m)n_0h^{d_m}} - \sum_{m_1\neq m_2\in\mathcal{M}} \frac{\gamma_{m_1}\gamma_{m_2}\tau_{m_1,m_2}(x^{m_{1,2}})\nu_{m_1,m_2}f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1}f_{X^{m_1}}(x^{m_1})\mu_{0,m_2}f_{X^{m_2}}(x^{m_2})n_0h^{d_{m^{1,2}}}};$$

$$R_{22} := \sigma^2\gamma^T H_{0,\mathcal{M}}^T H_{0,\mathcal{M}}\gamma - \sum_{m\in\mathcal{M}} \frac{\gamma_m^2\nu_m\sigma^2}{\mu_{0,m}^2 f_{X^m}(x^m)n_0h^{d_m}} - \sum_{m_1\neq m_2\in\mathcal{M}} \frac{\gamma_{m_1}\gamma_{m_2}\sigma^2\nu_{m_1,m_2}f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1}f_{X^{m_1}}(x^{m_1})\mu_{0,m_2}f_{X^{m_2}}(x^{m_2})n_0h^{d_{m^{1,2}}}};$$

and

$$R_{23} := \gamma^T H_{\mathcal{M}}^T(\sigma^2 I + \Gamma^*)H_{\mathcal{M}}\gamma - \sum_{m\in\mathcal{M}} \frac{\gamma_m^2\nu_m\{\sigma^2+\tau_m(x^m)\}}{\mu_{0,m}^2 f_{X^m}(x^m)n_mh^{d_m}}.$$

Note that $R_2 = R_{21} + R_{22} + R_{23}$, since

$$\gamma^T\Lambda_{\mathrm{R}}\gamma = \sum_{m\in\mathcal{M}} \frac{\gamma_m^2\nu_m\{\sigma^2+\tau_m(x^m)\}}{\mu_{0,m}^2 f_{X^m}(x^m)n_0h^{d_m}} + \sum_{m\in\mathcal{M}} \frac{\gamma_m^2\nu_m\{\sigma^2+\tau_m(x^m)\}}{\mu_{0,m}^2 f_{X^m}(x^m)n_mh^{d_m}} - \sum_{m_1\neq m_2\in\mathcal{M}} \frac{\gamma_{m_1}\gamma_{m_2}\{\sigma^2+\tau_{m_1,m_2}(x^{m_{1,2}})\}\nu_{m_1,m_2}f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1}f_{X^{m_1}}(x^{m_1})\mu_{0,m_2}f_{X^{m_2}}(x^{m_2})n_0h^{d_{m^{1,2}}}}.$$

We now bound R_{21}, R_{22} and R_{23} in turn.

To bound R_{21}: Write $E_m = (\eta_m(X_1^m), \dots \eta_m(X_n^m))^T$ and let $\delta_m = E - E_m$. Then $H_m\delta_m = 0$ and

$$\begin{aligned}\gamma^T\big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big)E &= \sum_{m\in\mathcal{M}} \gamma_m\big(H_{0,m}^T - H_m^T\big)E \\ &= \sum_{m\in\mathcal{M}} \gamma_m\big(H_{0,m}^T - H_m^T\big)(E - E_m + E_m) \\ &= \sum_{m\in\mathcal{M}} \gamma_m\big\{H_{0,m}^T\delta_m + (H_{0,m}^T - H_m^T)E_m\big\}. \end{aligned} \tag{29}$$

For $m \in \mathcal{M}$, using the fact $H_{0,m}^T 1_n = H_m^T 1_n = 1$ we have that

$$
\begin{aligned}
&|(H_{0,m}^T - H_m^T)E_m| = |(H_{0,m}^T - H_m^T)\{E_m - \eta_m(x^m)1_n\}| \\
&= \Big| \frac{\frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta_m(X_i^m) - \eta_m(x^m)\}}{\frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)} \\
&\qquad\qquad - \frac{\frac{1}{n_m h^{d_m}} \sum_{i \in A_m} K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta_m(X_i^m) - \eta_m(x^m)\}}{\frac{1}{n_m h^{d_m}} \sum_{i \in A_m} K_m\big(\frac{X_i^m - x^m}{h}\big)} \Big| \\
&= \frac{1}{\hat{f}_{0,m}\hat{f}_m n_0 n_m h^{2d_m}} \Big| \sum_{i \in A_0, j \in A_m} \Big[K_m\Big(\frac{X_i^m - x^m}{h}\Big)\{\eta_m(X_i^m) - \eta_m(x^m)\} K_m\Big(\frac{X_j^m - x^m}{h}\Big) \\
&\qquad\qquad - K_m\Big(\frac{X_j^m - x^m}{h}\Big)\{\eta_m(X_j^m) - \eta_m(x^m)\} K_m\Big(\frac{X_i^m - x^m}{h}\Big) \Big] \Big|,
\end{aligned}
$$

The last line in the display above is the sum of $n_0 n_m$ mean zero terms. Moreover, for each term we have

$$
\begin{aligned}
&\mathbb{E}\Big[K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta_m(X_i^m) - \eta_m(x^m)\} K_m\big(\frac{X_j^m - x^m}{h}\big) \\
&\qquad\qquad - K_m\big(\frac{X_j^m - x^m}{h}\big)\{\eta_m(X_j^m) - \eta_m(x^m)\} K_m\big(\frac{X_i^m - x^m}{h}\big) \Big]^2 \\
&= \int_{\mathbb{R}^{d_m}} \int_{\mathbb{R}^{d_m}} \Big[K_m\big(\frac{z_1 - x^m}{h}\big)\{\eta_m(z_1) - \eta_m(x^m)\} K_m\big(\frac{z_2 - x^m}{h}\big) \\
&\qquad\qquad - K_m\big(\frac{z_2 - x^m}{h}\big)\{\eta_m(z_2) - \eta_m(x^m)\} K_m\big(\frac{z_1 - x^m}{h}\big) \Big]^2 f_{X^m}(z_1^m) f_{X^m}(z_2^m)\, dz_1 dz_2 \\
&= 2 \int_{\mathbb{R}^{d_m}} \int_{\mathbb{R}^{d_m}} \Big[K_m^2\big(\frac{z_1 - x^m}{h}\big)\{\eta_m(z_1) - \eta_m(x^m)\}^2 K_m^2\big(\frac{z_2 - x^m}{h}\big) f_{X^m}(z_1^m) f_{X^m}(z_2^m)\, dz_1 dz_2 \\
&\qquad\qquad - 2 \int_{\mathbb{R}^{d_m}} \int_{\mathbb{R}^{d_m}} K_m^2\big(\frac{z_2 - x^m}{h}\big)\{\eta_m(z_1) - \eta_m(x^m)\} \\
&\qquad\qquad\qquad \{\eta_m(z_2) - \eta_m(x^m)\} K_m^2\big(\frac{z_1 - x^m}{h}\big) f_{X^m}(z_1^m) f_{X^m}(z_2^m)\, dz_1 dz_2 \\
&\le 2L^2 h^{2d_m+2} \int_{\mathbb{R}^{d_m}} K_m^2(u)\|u\|^2 f_{X^m}(x^m + hu)\, du \int_{\mathbb{R}^{d_m}} K_m^2(u) f_{X^m}(x^m + hu)\, dz_u \\
&\qquad\qquad + 2L^2 h^{2d_m+2} \Big\{ \int_{\mathbb{R}^{d_m}} K_m(u)\|u\| f_{X^m}(x^m + hu)\, du \Big\}^2 \\
&\le 2L^2 h^{2d_m+2} \{ f_{X^m}(x^m)\mu_{1,m}\bar{K}_m + Lh\mu_{2,m}\bar{K}_m + (f_{X^m}(x^m)\mu_{1,m} + Lh\mu_{1,m}\bar{K}_m)^2 \}.
\end{aligned}
$$

Therefore, by Markov's inequality we obtain that

$$
\begin{aligned}
&\frac{1}{n_0 n_m h^{2d_m}} \Big| \sum_{i \in A_0, j \in A_m} \Big[K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta_m(X_i^m) - \eta_m(x^m)\} K_m\big(\frac{X_j^m - x^m}{h}\big) \\
&\qquad - K_m\big(\frac{X_j^m - x^m}{h}\big)\{\eta_m(X_j^m) - \eta_m(x^m)\} K_m\big(\frac{X_i^m - x^m}{h}\big) \Big] \Big| = O_p\Big(\frac{1}{n_0^{1/2} n_m^{1/2} h^{d_m - 1}}\Big),
\end{aligned}
$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Using also the fact that $\hat{f}_{0,m} = f_{X^m}(x^m) + O_p(h)$ and $\hat{f}_m = f_{X^m}(x^m) + O_p(h)$, we have by Slutsky's Theorem that

$$|(H_{0,m}^T - H_m^T)E_m| = O_p\Big(\frac{1}{n_0^{1/2} n_m^{1/2} h^{d_m - 1}}\Big), \tag{30}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Now, for the terms involving δ_m in (29), first write

$$\begin{aligned} H_{0,m}^T \delta_m &= \frac{\frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta(X_i) - \eta_m(X_i^m)\}}{\frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)} \\ &= \frac{1}{\hat{f}_{0,m} n_0 h^{d_m}} \sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)\{\eta(X_i) - \eta_m(X_i^m)\}. \end{aligned}$$

The last term is a sum of n_0 mean zero terms, with

$$\begin{aligned} \mathbb{E}\Big[K_m^2\big(\frac{X_i^m - x^m}{h}\big)\{\eta(X_i) - \eta_m(X_i^m)\}^2\Big] &= \mathbb{E}\Big\{K_m^2\big(\frac{X_i^m - x^m}{h}\big)\tau_m(X_i^m)\Big\} \\ &= \int_{\mathbb{R}^{d_m}} K_m^2\big(\frac{z^m - x^m}{h}\big)\tau_m(z^m) f_{X^m}(z^m)\, dz \\ &= h^{d_m} \int_{\mathbb{R}^{d_m}} K_m^2(u)\tau_m(x^m + hu) f_{X^m}(x^m + hu)\, du \\ &\le h^{d_m}[\tau_m(x^m) f_{X^m}(x^m) + Lh\bar{K}_m \mu_{0,m}\{\tau_m(x^m) + f_{X^m}(x^m)\} + L^2 h^2 \bar{K}_m \mu_{1,m}]. \end{aligned}$$

Thus by Markov's inequality we have $|H_{0,m}^T \delta_m| = O_p(n_0^{-1/2} h^{-d_m/2})$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$, and it follows immediately that $|(H_{0,m}^T - H_m^T)E_m| = O_p\Big(h(H_{0,m}^T \delta_m)^2\Big)$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Thus, the first term in R_{21} can be written as

$$\{\gamma^T\big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big)E\}^2 = \Big(\sum_{m \in \mathcal{M}} \gamma_m H_{0,m}^T \delta_m\Big)^2 \{1 + O_p(h)\},$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Now observe that

$$\begin{aligned} &\mathbb{E}\{\sum_{m \in \mathcal{M}} \gamma_m^2 (H_{0,m}^T \delta_m)^2 | X_1^m, \ldots, X_n^m\} - \sum_{m \in \mathcal{M}} \frac{\gamma_m^2 \nu_m \tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m) n_0 h^{d_m}} \\ &= \sum_{m \in \mathcal{M}} \gamma_m^2 H_{0,m}^T \mathbb{E}\{\delta_m \delta_m^T | X_1^m, \ldots, X_n^m\} H_{0,m} - \sum_{m \in \mathcal{M}} \frac{\gamma_m^2 \nu_m \tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m) n_0 h^{d_m}} \\ &= \sum_{m \in \mathcal{M}} \gamma_m^2 H_{0,m}^T \mathrm{diag}\{\tau_m(X_1^m), \ldots, \tau_m(X_n^m)\} H_{0,m} - \sum_{m \in \mathcal{M}} \frac{\gamma_m^2 \nu_m \tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m) n_0 h^{d_m}} \\ &= \sum_{m \in \mathcal{M}} \gamma_m^2 H_{0,m}^T \mathrm{diag}\{\tau_m(X_1^m), \ldots, \tau_m(X_n^m)\} H_{0,m} - \sum_{m \in \mathcal{M}} \gamma_m^2 \tau_m(x^m) H_{0,m}^T H_{0,m} \\ &\qquad + \sum_{m \in \mathcal{M}} \gamma_m^2 \tau_m(x^m) H_{0,m}^T H_{0,m} - \sum_{m \in \mathcal{M}} \frac{\gamma_m^2 \nu_m \tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m) n_0 h^{d_m}}. \end{aligned}$$

Then, by assumption **A4**, we have that

$$\begin{aligned}
&\Big|\sum_{m\in\mathcal{M}}\gamma_m^2 H_{0,m}^T \mathrm{diag}\{\tau_m(X_1^m),\ldots,\tau_m(X_n^m)\}H_{0,m} - \sum_{m\in\mathcal{M}}\gamma_m^2\tau_m(x^m)H_{0,m}^T H_{0,m}\Big| \\
&= \Big|\sum_{m\in\mathcal{M}}\gamma_m^2 H_{0,m}^T \mathrm{diag}\{\tau_m(X_i^m)-\tau_m(x^m)\}H_{0,m}\Big| \\
&\le \sum_{m\in\mathcal{M}}\gamma_m^2 \frac{\sum_{i\in A_0} K_m^2\big(\frac{X_i^m-x^m}{h}\big)|\tau_m(X_i^m)-\tau_m(x^m)|}{\{\sum_{i\in A_0}K_m\big(\frac{X_i^m-x^m}{h}\big)\}^2} \\
&\le L\sum_{m\in\mathcal{M}}\gamma_m^2 \frac{\frac{1}{n_0^2h^{2d_m}}\sum_{i\in A_0}K_m^2\big(\frac{X_i^m-x^m}{h}\big)\|X_i^m-x^m\|}{\hat{f}_{0,m}} = O_p\Big(\sum_{m\in\mathcal{M}}\gamma_m^2\frac{1}{n_0h^{d_m-1}}\Big), \qquad (31)
\end{aligned}$$

uniformly for $h\in[n^{-\beta},n^{-\alpha}]$. Moreover, by (24), we have

$$\Big|\sum_{m\in\mathcal{M}}\gamma_m^2\tau_m(x^m)H_{0,m}^T H_{0,m} - \sum_{m\in\mathcal{M}}\frac{\gamma_m^2\nu_m\tau_m(x^m)}{\mu_{0,m}^2 f_{X^m}(x^m)n_0h^{d_m}}\Big| = O_p\Big(\sum_{m\in\mathcal{M}}\gamma_m^2\frac{1}{n_0h^{d_m-1}}\Big), \qquad (32)$$

uniformly for $h\in[n^{-\beta},n^{-\alpha}]$. Similarly

$$\begin{aligned}
&\mathbb{E}\Big\{\sum_{m_1\neq m_2\in\mathcal{M}}\gamma_{m_1}\gamma_{m_2}H_{0,m_1}^T\delta_{m_1}H_{0,m_2}^T\delta_{m_2}\Big|X_1^{m_{1,2}},\ldots,X_n^{m_{1,2}}\Big\} \\
&\qquad - \sum_{m_1\neq m_2\in\mathcal{M}}\frac{\gamma_{m_1}\gamma_{m_2}\tau_{m_1,m_2}(x^{m_{1,2}})\nu_{m_1,m_2}f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1}f_{X^{m_1}}(x^{m_1})\mu_{0,m_2}f_{X^{m_2}}(x^{m_2})n_0h^{d_{m^{1,2}}}}; \\
&= \sum_{m_1\neq m_2\in\mathcal{M}}\gamma_{m_1}\gamma_{m_2}H_{0,m_1}^T\Big[\mathbb{E}\Big\{\delta_{m_1}\delta_{m_2}^T\Big|X_1^{m_{1,2}},\ldots,X_n^{m_{1,2}}\Big\}-\tau_{m_1,m_2}(x^{m_{1,2}})I_{n\times n}\Big]H_{0,m_2} \\
&\qquad + \sum_{m_1\neq m_2\in\mathcal{M}}\gamma_{m_1}\gamma_{m_2}H_{0,m_1}^T H_{0,m_2}\tau_{m_1,m_2}(x^{m_{1,2}}) \\
&\qquad - \sum_{m_1\neq m_2\in\mathcal{M}}\frac{\gamma_{m_1}\gamma_{m_2}\tau_{m_1,m_2}(x^{m_{1,2}})\nu_{m_1,m_2}f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1}f_{X^{m_1}}(x^{m_1})\mu_{0,m_2}f_{X^{m_2}}(x^{m_2})n_0h^{d_{m^{1,2}}}}.
\end{aligned}$$

Now, by the last part of Assumption **A4** and similar arguments to those used above, we have

$$\begin{aligned}
&\Big|\sum_{m_1\neq m_2\in\mathcal{M}}H_{0,m_1}^T\Big[\mathbb{E}\Big\{\delta_{m_1}\delta_{m_2}^T\Big|X_1^{m_{1,2}},\ldots,X_n^{m_{1,2}}\Big\}-\tau_{m_1,m_2}(x^{m_{1,2}})I_{n\times n}\Big]H_{0,m_2}\Big| \\
&\qquad = O_p\Big(\sum_{m_1\neq m_2\in\mathcal{M}}\gamma_{m_1}\gamma_{m_2}\frac{1}{n_0h^{d_{m^{1,2}}-1}}\Big), \qquad (33)
\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Furthermore, by (26), we have

$$\Big| \sum_{m_1 \neq m_2 \in \mathcal{M}} \gamma_{m_1}\gamma_{m_2} H_{0,m_1}^T H_{0,m_2} \tau_{m_1,m_2}(x^{m_{1,2}}) - \sum_{m_1 \neq m_2 \in \mathcal{M}} \frac{\gamma_{m_1}\gamma_{m_2}\tau_{m_1,m_2}(x^{m_{1,2}})\nu_{m_1,m_2} f_{X^{m_{1,2}}}(x^{m_{1,2}})}{\mu_{0,m_1} f_{X^{m_1}}(x^{m_1}) \mu_{0,m_2} f_{X^{m_2}}(x^{m_2}) n_0 h^{d_{m^{1,2}}}} \Big| = O_p\Big(\sum_{m_1 \neq m_2 \in \mathcal{M}} \frac{1}{n_0 h^{d_{m^{1,2}}-1}} \Big), \tag{34}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Then, using (30), (31), (32), (33) and (34), we conclude that

$$|R_{21}| = O_p\Big(\sum_{m \in \mathcal{M}} \frac{1}{n_0 h^{d_m - 1}} + \sum_{m_1 \neq m_2 \in \mathcal{M}} \frac{1}{n_0 h^{d_{m^{1,2}}-1}} \Big),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

Furthermore, by (24) and (26), we have that

$$|R_{22}| = O_p\Big(\sum_{m \in \mathcal{M}} \frac{1}{n_0 h^{d_m - 1}} + \sum_{m_1 \neq m_2 \in \mathcal{M}} \frac{1}{n_0 h^{d_{m^{1,2}}-1}} \Big),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. Finally, by (25) we have

$$|R_{23}| = O_p\Big(\sum_{m \in \mathcal{M}} \frac{1}{n_m h^{d_m - 1}} \Big),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. This concludes the bound on R_2 and Part I of the proof.

Part II: the second term in (21). Using the definition of local linear estimators and noting that $\Gamma H_0 = 0$ and $H_{\mathcal{M}}^T H_0 = 0$, the second term in (21) can be written as

$$\begin{aligned}
&\mathbb{E}[(\hat{\eta}_\gamma^{\mathcal{M}} - \hat{\eta}_0)\{\hat{\eta}_0 - \eta(x)\} | X_1^{m_1}, \ldots, X_n^{m_n}] \\
&= \mathbb{E}\Big[(\hat{\eta}_\gamma^{\mathcal{M}} - \hat{\eta}_0)\big\{\hat{\eta}_0 - \mathbb{E}(\hat{\eta}_0 | X_1^{m_1}, \ldots, X_n^{m_n}) + \mathbb{E}(\hat{\eta}_0 | X_1^{m_1}, \ldots, X_n^{m_n}) - \eta(x)\big\} \big| X_1^{m_1}, \ldots, X_n^{m_n} \Big] \\
&= -\gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big) \mathbb{E}\{Y_{[n]}(Y_{[n]}^T - E^T) | X_1^{m_1}, \ldots, X_n^{m_n}\} H_0 \\
&\qquad\qquad - \gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big) E\{E^T H_0 - \eta(x)\} \\
&= -\gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big)(\sigma^2 I + \Gamma) H_0 - \gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big) E\{E^T H_0 - \eta(x)\} \\
&= -\sigma^2 \gamma^T H_{0,\mathcal{M}}^T H_0 - \gamma^T \big(H_{0,\mathcal{M}}^T - H_{\mathcal{M}}^T\big) E\{E^T H_0 - \eta(x)\}.
\end{aligned} \tag{35}$$

For the first term in (35), by (23) we have

$$\sigma^2 \gamma^T H_{0,\mathcal{M}}^T H_0 = \sigma^2 \gamma^T \Omega_{\mathrm{R}} \{1 + O_p(h)\},$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

It remains to bound the second term. Let

$$R_3 = -\gamma^T (H_{0,\mathcal{M}} - H_{\mathcal{M}}) E\{E^T H_0 - \eta(x)\}.$$

Recall from Part I of this proof that $\gamma^T(H_{0,\mathcal{M}} - H_{\mathcal{M}})E = O_p\big(\sum_{m\in\mathcal{M}} \frac{\gamma_m}{n_0 h^{d_m}}\big)$. Moreover, by assumption **A3**, we have that

$$|\eta(X_i) - \eta(x)| \leq L\|X_i - x\|,$$

for $i = 1, \ldots, n$. Recall that $E^{*T}H_0 = \eta(x)$. Therefore

$$\begin{aligned}|E^T H_0 - \eta(x)| = |(E^T - E^{*T})H_0| &\leq \frac{\sum_{i\in A_0} K\big(\frac{X_i - x}{h}\big)|\eta(X) - \eta(x)|}{\sum_{i\in A_0} K\big(\frac{X_i - x}{h}\big)} \\ &\leq \frac{1}{\hat{f}_0 n_0 h^d} \sum_{i\in A_0} K\big(\frac{X_i - x}{h}\big)|\eta(X) - \eta(x)| = O_p(h),\end{aligned}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$, where the last equality is by Markov's inequality due to the facts that $K\big(\frac{X_i - x}{h}\big)|\eta(X) - \eta(x)| > 0$ and

$$\mathbb{E}\big\{K\big(\frac{X_i - x}{h}\big)|\eta(X) - \eta(x)|\big\} \leq Lh^{d+1}\int_{\mathbb{R}^d} K(u)\|u\| f_X(x + hu)\, du \leq Lh^{d+1}\{\mu_1 f_X(x) + Lh\mu_2\}.$$

This completes the proof.

■

A.5 Proofs for the results in Section 5

Proof of Corollary 7. Consider the bias and variance separately: we claim

$$\mathbb{E}(\hat{\varphi}^*_m) = \mathbb{E}(\hat{\theta}_0) \tag{36}$$

and

$$\mathrm{Var}(\hat{\varphi}^*_m) = \mathrm{Var}(\hat{\theta}_0) - \frac{\mathrm{Cov}(\hat{\theta}_0, \bar{\varphi}_{A_0,m})^2}{\mathrm{Var}(\bar{\varphi}_{A_0,m} - \bar{\varphi}_{\bar{A}_m,m})} \leq \mathrm{Var}(\hat{\theta}_0). \tag{37}$$

To see (36) it suffices to show that $\mathbb{E}(\bar{\varphi}_{A_0,m}) = \mathbb{E}(\bar{\varphi}_{\bar{A}_m,m})$. First, since the missing data is MCAR, we have that the data in $\mathcal{T}_{A_0,m} \cup \mathcal{T}_{\bar{A}_m,m}$ are independent and identically distributed with distribution Q_m. Furthermore, for each b_1, b_2, the estimators $\hat{\varphi}_{\tilde{A}_0^{b_1},m}$ and $\hat{\varphi}_{\tilde{A}_m^{b_2},m}$ are constructed using $n_{0,m}$ pairs from $\mathcal{T}_{A_0,m}$ and $\mathcal{T}_{\bar{A}_m,m}$, respectively. Thus, these two estimators have the same distribution, and in particular, they have the same mean. It follows that $\mathbb{E}(\bar{\varphi}_{A_0,m}) = \mathbb{E}(\bar{\varphi}_{\bar{A}_m,m})$. The result in (37) follows via a direct calculation. ■

A.6 Auxilliary Lemmas

The results in this section motivate our choice of the d_m-dimensional kernel, K_m, used to construct $\hat{f}_{0,m}$ and $\hat{f}_m$ in the density estimation problem, and $\hat{\eta}_{0,m}$ and $\hat{\eta}_m$ in the regression problem. For $m \in \{0,1\}^d$, let $m^c = (1, \ldots, 1)^T - m$. Recall that we write $f_{X|X^m}(x^{m^c}; x^m)$, for the conditional density of X given $X^m = x^m$ at x^{m^c}.

We see from Lemma 8 that, up to rescaling, if the kernel can be factorised as $K(t) = K_m(t^m)K_{m^c}(t^{m^c})$, then the optimal choice of $\hat{f}_{0,m}$ is well approximated by

$$\frac{1}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\Big(\frac{X_i^m - x^m}{h}\Big).$$

On the other hand, Lemma 9 shows that the optimal choice of $\hat{\eta}_{0,m}$ is well approximated by

$$\frac{\sum_{i \in A_0} Y_i K_m\big(\frac{X_i^m - x^m}{h}\big)}{\sum_{i \in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)}.$$

Any d-dimensional kernel constructed from the product of 1-dimensional kernels satisfy the factor assumption in Lemmas 8 and 9 . For example the condition is satisfied by the Gaussian kernel $\frac{1}{(2\pi)^{d/2}} \exp(-\|t\|^2/2)$ and the box kernel $\frac{1}{2^d} \mathbb{1}_{\{\max_{j=1,\ldots,d} |t_j| \le 1\}}$. Moreover, other kernels, such as $\mathbb{1}_{\{\|t\|^2 \le 1\}}$, are not covered by the lemma but enjoy similar properties to that in (38), which can be proved by using similar ideas.

Lemma 8 *Assume **A1** and suppose that, for $t \in \mathbb{R}^d$ and $m \in \{0,1\}^d$, we have $K(t) = K_m(t^m)K_{m^c}(t^{m^c})$, for some $K_m : \mathbb{R}^{d_m} \to [0,\infty)$ and $K_{m^c} : \mathbb{R}^{d-d_m} \to [0,\infty)$, that satisfy $\mu_{0,m} = \int_{\mathbb{R}^{d_m}} K_m(z)\,dz < \infty$, $\mu_{1,m} = \int_{\mathbb{R}^{d_m}} \|z\| K_m(z)\,dz < \infty$, $\mu_{0,m^c} = \int_{\mathbb{R}^{d_{m^c}}} K_{m^c}(z)\,dz < \infty$ and $\mu_{1,m^c} = \int_{\mathbb{R}^{d_{m^c}}} \|z\| K_{m^c}(z)\,dz < \infty$. Then, for $z \in \mathbb{R}^{d_m}$,*

$$\Big|\mathbb{E}\Big\{K\Big(\frac{X-x}{h}\Big)\Big|X^m = z^m\Big\} - h^{d-d_m} K_m\Big(\frac{z^m - x^m}{h}\Big) f_{X|X^m}(x^{m^c}; z^m)\mu_{0,m^c}\Big| \le \frac{L h^{1+d-d_m}}{f_{X^m}(z^m)} K_m\Big(\frac{z^m - x^m}{h}\Big)\mu_{1,m^c}. \tag{38}$$

Therefore, for $0 < \alpha < \beta < 1/d$,

$$\frac{1}{n_0 h^d} \sum_{i \in A_0} \mathbb{E}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big|X_i^m\Big\} = \frac{\mu_{0,m^c}}{n_0 h^{d_m}} \sum_{i \in A_0} K_m\Big(\frac{X_i^m - x^m}{h}\Big) f_{X|X^m}(x^{m^c}; x^m) + O_p(h)$$

as $n \to \infty$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

Proof To see (38), first observe that by making the substitution $u = \frac{z^{m^c} - x^{m^c}}{h}$, we have

$$\begin{aligned}
&\mathbb{E}\Big\{K\Big(\frac{X-x}{h}\Big)\Big|X^m = z^m\Big\} \\
&\quad = \int_{\mathbb{R}^{d-d_m}} K_m\Big(\frac{z^m - x^m}{h}\Big) K_{m^c}\Big(\frac{z^{m^c} - x^{m^c}}{h}\Big) f_{X|X^m}(z^{m^c}; z^m)\,dz^{m^c} \\
&\quad = h^{d-d_m} K_m\Big(\frac{z^m - x^m}{h}\Big) \int_{\mathbb{R}^{d-d_m}} K_{m^c}(u) f_{X|X^m}(x^{m^c} + hu; z^m)\,du.
\end{aligned}$$

Now, write

$$f_{X|X^m}(z^{m^c}; z^m) = \frac{f_X(z)}{f_{X^m}(z^m)}.$$

Thus, by assumption **A1**, we have that

$$|f_{X|X^m}(x^{m^c}+hu;z^m)-f_{X|X^m}(x^{m^c};z^m)|\le\frac{Lh\|u\|}{f_{X^m}(z^m)}.$$

It follows that

$$\Big|\mathbb{E}\Big\{K\Big(\frac{X-x}{h}\Big)\Big|X^m=z^m\Big\}-h^{d-d_m}K_m\Big(\frac{z^m-x^m}{h}\Big)f_{X|X^m}(x^{m^c};z^m)\mu_{0,m^c}\Big|$$
$$\le\frac{Lh^{1+d-d_m}}{f_{X^m}(z^m)}K_m\Big(\frac{z^m-x^m}{h}\Big)\int_{\mathbb{R}^{d-d_m}}K_{m^c}(u)\|u\|\,du.$$

This proves (38).

For the remainder of the proof, first observe that

$$\mathbb{E}\Big\{\frac{1}{f_{X^m}(X_i^m)}K_m\Big(\frac{X_i^m-x^m}{h}\Big)\Big\}=\int_{\mathbb{R}^{d_m}}K_m\Big(\frac{z^m-x^m}{h}\Big)\,dz=h^{d_m}\mu_{0,m}.$$

It remains to bound the following:

$$\begin{aligned}
&\mathbb{E}\Big|\frac{1}{n_0h^{d_m}}\sum_{i\in A_0}K_m\Big(\frac{X_i^m-x^m}{h}\Big)\{f_{X|X^m}(x^{m^c};X_i^m)-f_{X|X^m}(x^{m^c};x^m)\}\Big|\\
&=\frac{1}{h^{d_m}}\int_{\mathbb{R}^{d_m}}\Big|K_m\Big(\frac{z^m-x^m}{h}\Big)\{f_{X|X^m}(x^{m^c};z^m)-f_{X|X^m}(x^{m^c};x^m)\}\Big|f_{X^m}(z^m)\,dz^m\\
&=\int_{\mathbb{R}^{d_m}}K_m(u)|f_{X|X^m}(x^{m^c};x^m+hu)-f_{X|X^m}(x^{m^c};x^m)|f_{X^m}(x^m+hu)\,dz^m\\
&=\Big|\int_{\mathbb{R}^{d_m}}K_m(u)\big|f_X(x^{m^c},x^m+hu)-\frac{f_X(x)f_{X^m}(x^m+hu)}{f_{X^m}(x^m)}\big|\,du\\
&\le Lh\Big\{1+\frac{f_X(x)}{f_{X^m}(x^m)}\Big\}\int_{\mathbb{R}^{d_m}}K_m(u)\|u\|\,du=Lh\Big\{1+\frac{f_X(x)}{f_{X^m}(x^m)}\Big\}\mu_{1,m}.
\end{aligned}$$

The proof is completed using Markov's inequality. ■

Our results on the choice of K_m in the regression problem need a slightly stronger condition on the joint distribution of (X,Y):

A5 *Suppose that P and P_m, for $m\in\{0,1\}^d\backslash\{(1,\dots,1)^T\}$, have densities $f_{X,Y}$, and $f_{X^m,Y}$, respectively, that satisfy $|f_{X,Y}(z_1,y)-f_{X,Y}(z_2,y)|\le L\|z_1-z_2\|$, for all $z_1,z_2\in\mathbb{R}^d,y\in\mathbb{R}$, and, for each $m\in\{0,1\}^d\setminus\{(1,\dots,1)^T\}$, we have $|f_{X^m,Y}(z_1^m,y)-f_{X^m,Y}(z_2^m,y)|\le L\|z_1^m-z_2^m\|$, for all $z_1^m,z_2^m\in\mathbb{R}^{d_m},y\in\mathbb{R}$. Moreover, we ask that the densities f_X, $f_{X^m,Y}$ are bounded, and that $f_{X^{m^c}|X^m,Y}(x^m;z^m,y)$ is a bounded function of z^m and y.*

Lemma 9 *Assume **A3** and **A5**, and suppose that Y is supported on $\mathcal{D}_Y\subseteq[-\bar{Y},\bar{Y}]$, for some $\bar{Y}>0$. Suppose further that for $t\in\mathbb{R}^d$ and $m\in\{0,1\}^d$, we have $K(t)=K_m(t^m)K_{m^c}(t^{m^c})$, for some $K_m:\mathbb{R}^{d_m}\to[0,\infty)$ and $K_{m^c}:\mathbb{R}^{d-d_m}\to[0,\infty)$, that*

satisfy $\bar{K}_m := \sup_{z\in\mathbb{R}^{d_m}}(1+\|z\|)K_m(z) < \infty$, $\mu_{0,m} = \int_{\mathbb{R}^{d_m}} K_m(z)\,dz < \infty$, $\mu_{1,m} = \int_{\mathbb{R}^{d_m}} \|z\|K_m(z)\,dz < \infty$, $\mu_{0,m^c} = \int_{\mathbb{R}^{d_{m^c}}} K_{m^c}(z)\,dz < \infty$ *and* $\mu_{1,m^c} = \int_{\mathbb{R}^{d_{m^c}}} \|z\|K_{m^c}(z)\,dz < \infty$. *Then, for* $0<\alpha<\beta<1/d$,

$$\sum_{i\in A_0} Y_i \mathbb{E}\Big\{ \frac{K\big(\frac{X_i-x}{h}\big)}{\sum_{j\in A_0} K\big(\frac{X_j-x}{h}\big)} \Big| \mathcal{T}_{A_0,m} \Big\} = \frac{\eta(x)}{\eta_m(x_m)} \frac{\sum_{i\in A_0} Y_i K_m\big(\frac{X_i^m - x^m}{h}\big)}{\sum_{j\in A_0} K_m\big(\frac{X_j^m - x^m}{h}\big)} + o_p(1)$$

as $n\to\infty$, *uniformly for* $h\in[n^{-\beta}, n^{-\alpha}]$.

Proof *Part I*: We first show that

$$\Pi_0 := \sum_{i\in A_0} Y_i \mathbb{E}\Big\{ \frac{K\big(\frac{X_i-x}{h}\big)}{\sum_{j\in A_0} K\big(\frac{X_j-x}{h}\big)} \Big| \mathcal{T}_{A_0,m} \Big\} - \sum_{i\in A_0} Y_i \frac{\mathbb{E}\big\{K\big(\frac{X_i-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}}{\mathbb{E}\big\{\sum_{j\in A_0} K\big(\frac{X_j-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}} = o_p(1). \tag{39}$$

To see (39), let

$$W_1 := \frac{1}{n_0 h^d} \sum_{i\in A_0} Y_i \Big[K\Big(\frac{X_i-x}{h}\Big) - \mathbb{E}\Big\{ K\Big(\frac{X_i-x}{h}\Big) \Big| X_i^m, Y_i \Big\}\Big]$$

and

$$W_2 := \frac{1}{n_0 h^d} \sum_{i\in A_0} \Big[K\Big(\frac{X_i-x}{h}\Big) - \mathbb{E}\Big\{ K\Big(\frac{X_i-x}{h}\Big) \Big| X_i^m, Y_i \Big\}\Big].$$

Then we can write

$$\hat{\eta}_0 = \frac{\sum_{i\in A_0} Y_i \mathbb{E}\big\{K\big(\frac{X_i-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}}{\mathbb{E}\big\{\sum_{j\in A_0} K\big(\frac{X_j-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}} + \frac{W_1}{\hat{f}_0} - \frac{W_2}{\hat{f}_0} \frac{\sum_{i\in A_0} Y_i \mathbb{E}\big\{K\big(\frac{X_i-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}}{\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})}$$

It follows that

$$\Pi_0 = \mathbb{E}\Big(\frac{W_1}{\hat{f}_0}\Big|\mathcal{T}_{A_0,m}\Big) + \mathbb{E}\Big(\frac{W_2}{\hat{f}_0}\Big|\mathcal{T}_{A_0,m}\Big) \frac{\frac{1}{n_0h^d}\sum_{i\in A_0} Y_i \mathbb{E}\big\{K\big(\frac{X_i-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}}{\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})}.$$

Now, by Markov's inequality, we have

$$\mathbb{P}(|W_1|>t|\mathcal{T}_{A_0,m}) \le \frac{\mathbb{E}(W_1^2|\mathcal{T}_{A_0,m})}{t^2} = \frac{1}{n_0^2h^{2d}t^2} \sum_{i\in A_0} Y_i^2 \mathrm{Var}\Big\{ K\Big(\frac{X_i-x}{h}\Big)\Big| X_i^m, Y_i\Big\}$$

and

$$\mathbb{P}(|W_2|>t|\mathcal{T}_{A_0,m}) \le \frac{\mathbb{E}(W_2^2|\mathcal{T}_{A_0,m})}{t^2} = \frac{1}{n_0^2h^{2d}t^2} \sum_{i\in A_0} \mathrm{Var}\Big\{ K\Big(\frac{X_i-x}{h}\Big)\Big| X_i^m, Y_i\Big\}.$$

Note further that, since $|Y_i|\le \bar{Y}$, for each i, we have that $|\Pi_0|\le 2\bar{Y}$. Therefore, for $0<t<\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})/2$,

$$\begin{aligned}
\big|\mathbb{E}\big(\Pi_0|\mathcal{T}_{A_0,m}\big)\big| &\le \mathbb{E}\big(|\Pi_0|\mathbb{1}_{\{|W_1|<t\}}\mathbb{1}_{\{|W_2|<t\}}|\mathcal{T}_{A_0,m}\big) + \mathbb{E}\big\{|\Pi_0|(\mathbb{1}_{\{|W_1|>t\}}+\mathbb{1}_{\{|W_2|>t\}})|\mathcal{T}_{A_0,m}\big\} \\
&\le \frac{2t}{\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})} + \frac{2t}{\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})} \frac{\frac{1}{n_0h^d}\sum_{i\in A_0} Y_i \mathbb{E}\big\{K\big(\frac{X_i-x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\}}{\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m})} \\
&\qquad + \frac{2\bar{Y}(1+\bar{Y}^2)}{n_0^2h^{2d}t^2} \sum_{i\in A_0} \mathrm{Var}\big\{K\big(\frac{X_i-x}{h}\big)\big|X_i^m, Y_i\big\}.
\end{aligned} \tag{40}$$

Now, using the tower property of expectation, we have

$$\mathbb{E}\big|\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m}) - f_X(x)\big| \le \mathbb{E}\big|\hat{f}_0 - f_X(x)\big| \le Lh\mu_1,$$

by assumption **A1**. Thus, we have $\mathbb{E}(\hat{f}_0|\mathcal{T}_{A_0,m}) = f_X(x) + O_p(h)$, by Markov's inequality. Similarly

$$\frac{1}{n_0 h^d}\sum_{i\in A_0} Y_i \mathbb{E}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big|\mathcal{T}_{A_0,m}\Big\} = \eta(x)f_X(x) + O_p(h).$$

Finally, using the facts that

$$\mathbb{E}\Big[\mathrm{Var}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big|X_i^m, Y_i\Big\}\Big] \le \mathrm{Var}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big\} \le h^d\{\nu f_X(x) + Lh\bar{K}\mu_0\},$$

and that the conditional variances are all non-negative, by applying Markov's inequality we have

$$\frac{1}{n_0}\sum_{i\in A_0} \mathrm{Var}\Big\{K\Big(\frac{X_i - x}{h}\Big)\Big|X_i^m, Y_i\Big\} = O_p(h^d),$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. The proof of Part I is then completed by taking $t = t_n := \sqrt{\frac{\log n_0}{n_0 h^d}}$ in (40) and using Markov's inequality.

Part II: We claim that

$$\Pi_1 := \frac{1}{n_0 h^d}\sum_{i\in A_0} Y_i\Big[\mathbb{E}\Big\{K\big(\frac{X_i - x}{h}\big)\Big|\mathcal{T}_{A_0,m}\Big\} - \frac{\mu_{0,m^c} f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)} K_m\big(\frac{X_i^m - x^m}{h}\big)\Big] = o_p(1), \tag{41}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$, and

$$\Pi_2 = \frac{1}{n_0 h^d}\sum_{i\in A_0} \mathbb{E}\Big\{K\big(\frac{X_i - x}{h}\big)\Big|\mathcal{T}_{A_0,m}\Big\} - \frac{\mu_{0,m^c} f_X(x)}{f_{X^m}(x^m) n_0 h^{d_m}} \sum_{i\in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big) = o_p(1), \tag{42}$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

To see (41), first observe that, for all $z^m \in \mathbb{R}^{d_m}$ and $y \in \mathbb{R}$, we have

$$\begin{aligned}
&\Big|\mathbb{E}\Big\{K_{m^c}\big(\frac{X_i^{m^c} - x^{m^c}}{h}\big)\Big|(X_i^m, Y_i) = (z^m, y)\Big\} - h^{d_{m^c}}\mu_{0,m^c} f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y)\Big| \\
&= \Big|\int_{\mathbb{R}^{d_{m^c}}} K_{m^c}\big(\frac{z^{m^c} - x^{m^c}}{h}\big) f_{X^{m^c}|X^m,Y}(z^{m^c}; z^m, y)\, dz^{m^c} - h^{d_{m^c}}\mu_{0,m^c} f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y)\Big| \\
&= h^{d_{m^c}}\Big|\int_{\mathbb{R}^{d_{m^c}}} K_{m^c}(u) f_{X^{m^c}|X^m,Y}(x^{m^c} + hu; z^m, y)\, du^{m^c} - \mu_{0,m^c} f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y)\Big| \\
&\le \frac{Lh^{d_{m^c}+1}\mu_{1,m^c}}{f_{X^m,Y}(z^m, y)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\mathbb{E}\Big|\frac{1}{n_0h^d}\sum_{i\in A_0}Y_i\Big[K_m\big(\frac{X_i^m-x^m}{h}\big)\mathbb{E}\big\{K_{m^c}\big(\frac{X_i^{m^c}-x^{m^c}}{h}\big)\big|X_i^m,Y_i\big\}\\
&\qquad\qquad - K_m\big(\frac{X_i^m-x^m}{h}\big)\mu_{0,m^c}f_{X^{m^c}|X^m,Y}(x^{m^c};X_i^m,Y_i)\Big]\Big|\\
&\le Lh\mu_{1,m^c}\mathbb{E}\Big\{\frac{1}{n_0h^{d_m}}\sum_{i\in A_0}\frac{|Y_i|K_m\big(\frac{X_i^m-x^m}{h}\big)}{f_{X^m,Y}(X_i^m,Y_i)}\Big\}\\
&= Lh\mu_{1,m^c}\frac{1}{h^{d_m}}\int_{\mathbb{R}^{d_m}\times\mathcal{D}_Y}|y|K_m\big(\frac{z^m-x^m}{h}\big)\,dzdy \le Lh\mu_{1,m^c}\mu_{0,m}\bar{Y}^2.
\end{aligned}$$

It follows that

$$\Pi_1=\frac{\mu_{0,m^c}}{n_0h^{d_m}}\sum_{i\in A_0}Y_iK_m\big(\frac{X_i^m-x^m}{h}\big)\Big[f_{X^{m^c}|X^m,Y}(x^{m^c};X_i^m,Y_i)-\frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big]+R_4, \tag{43}$$

where $R_4=O_p(h)$, uniformly for $h\in[n^{-\beta},n^{-\alpha}]$, by Markov's inequality. Now, writing (with a slight abuse of notation) $\eta(z^{m^c},z^m):=\mathbb{E}(Y|X^{m^c}=z^{m^c},X^m=z^m)$, the first term in (43) is the average of n_0 independent and identically distributed terms each with expectation

$$\begin{aligned}
&\frac{1}{h^{d_m}}\int_{\mathbb{R}^{d+1}}yK_m\big(\frac{z^m-x^m}{h}\big)\Big\{f_{X^{m^c}|X^m,Y}(x^{m^c};z^m,y)-\frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}f_{X^m,Y}(z^m,y)\,dz^mdy\\
&=\frac{1}{h^{d_m}}\int_{\mathbb{R}^{d+1}}yK_m\big(\frac{z^m-x^m}{h}\big)\Big\{f_{X^{m^c},X^m,Y}(x^{m^c},z^m,y)-\frac{f_X(x)\eta(x)f_{X^m,Y}(z^m,y)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}\,dz^mdy\\
&=\frac{1}{h^{d_m}}\int_{\mathbb{R}^d}K_m\big(\frac{z^m-x^m}{h}\big)\Big\{\eta(x^{m^c},z^m)f_{X^{m^c},X^m}(x^{m^c},z^m)-\frac{f_X(x)\eta(x)\eta_m(z^m)f_{X^m}(z^m)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}\,dz^m\\
&=\int_{\mathbb{R}^d}K_m(u)\Big\{\eta(x^{m^c},x^m+hu)f_{X^{m^c},X^m}(x^{m^c},x^m+hu)\\
&\qquad\qquad-\frac{f_X(x)\eta(x)\eta_m(x^m+hu)f_{X^m}(x^m+hu)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}\,du\\
&=\int_{\mathbb{R}^d}K_m(u)\Big\{\eta(x^{m^c},x^m+hu)f_{X^{m^c},X^m}(x^{m^c},x^m+hu)-f_X(x)\eta(x)\Big\}\,du\\
&\qquad+\mu_{0,m^c}\int_{\mathbb{R}^d}K_m(u)\Big\{f_X(x)\eta(x)-\frac{f_X(x)\eta(x)\eta_m(x^m+hu)f_{X^m}(x^m+hu)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}\,du\\
&\le Lh\Big[\mu_{1,m}\big\{\eta(x)+f_X(x)+\frac{f_{X^m}(x^m)+\eta_m(x^m)}{f_{X^m}(x^m)\eta_m(x^m)}\big\}+Lh\mu_{1,m}\bar{K}_m\big\{1+\frac{1}{f_{X^m}(x^m)\eta_m(x^m)}\big\}\Big].
\end{aligned}$$

Furthermore, we show below that

$$\mathrm{Var}\Big\{\frac{\mu_{0,m^c}}{h^{d_m}}Y_iK_m\big(\frac{X_i^m-x^m}{h}\big)\Big(f_{X^{m^c}|X^m,Y}(x^{m^c};X_i^m,Y_i)-\frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big)\Big\}=O\Big(\frac{1}{h^{d_m}}\Big), \tag{44}$$

uniformly for $h\in[n^{-\beta},n^{-\alpha}]$. It follows that $|\Pi_1|=o_p(1)$, uniformly for $h\in[n^{-\beta},n^{-\alpha}]$, by Chebychev's inequality – this proves (41).

Next, to see (42), write

$$\begin{aligned}\Pi_2 &= \frac{1}{n_0}\sum_{i\in A_0}\Big[\frac{1}{h^d}\mathbb{E}\big\{K\big(\frac{X_i - x}{h}\big)\big|\mathcal{T}_{A_0,m}\big\} - \frac{\mu_{0,m^c} f_X(x)}{f_{X^m}(x^m)h^{d_m}}K_m\big(\frac{X_i^m - x^m}{h}\big)\Big] \\ &= \frac{1}{n_0}\sum_{i\in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)\Big[\frac{1}{h^d}\mathbb{E}\Big\{K_{m^c}\big(\frac{X_i^{m^c} - x^{m^c}}{h}\big)\Big|X_i^m, Y_i\Big\} - \frac{\mu_{0,m^c} f_X(x)}{f_{X^m}(x^m)h^{d_m}}\Big] \\ &= \frac{\mu_{0,m^c}}{n_0 h^{d_m}}\sum_{i\in A_0} K_m\big(\frac{X_i^m - x^m}{h}\big)\Big[f_{X^{m^c}|X^m,Y}(x^{m^c}; X_i^m, Y_i) - \frac{f_X(x)}{f_{X^m}(x^m)}\Big] + R_5, \qquad (45)\end{aligned}$$

where, using the same technique used to bound R_4, we have

$$\begin{aligned}\mathbb{E}|R_5| &\le Lh\mu_{1,m^c}\mathbb{E}\Big\{\frac{1}{n_0 h^{d_m}}\sum_{i\in A_0}\frac{K_m\big(\frac{X_i^m - x^m}{h}\big)}{f_{X^m,Y}(X_i^m, Y_i)}\Big\} \\ &= Lh\mu_{1,m^c}\frac{1}{h^{d_m}}\int_{\mathbb{R}^{d_m}\times\mathcal{D}_Y} K_m\big(\frac{z^m - x^m}{h}\big)\,dzdy \le 2Lh\mu_{1,m^c}\mu_{0,m}\bar{Y}.\end{aligned}$$

Thus by Markov's inequality $R_5 = O_p(h)$, uniformly for $h \in [n^\beta, n^{-\alpha}]$. Finally, the remaining term in (45) is the average of n_0 independent and identically distributed terms, each with expectation given by

$$\begin{aligned}&\frac{1}{h^{d_m}}\int_{\mathbb{R}^d\times\mathbb{R}} K_m\big(\frac{z^m - x^m}{h}\big)\Big\{f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y) - \frac{f_X(x)}{f_{X^m}(x^m)}\Big\}f_{X^m,Y}(z^m, y)\,dz^m dy \\ &= \frac{\mu_{0,m^c}}{h^{d_m}}\int_{\mathbb{R}^d\times\mathbb{R}} K_m\big(\frac{z^m - x^m}{h}\big)\Big\{f_{X^{m^c},X^m,Y}(x^{m^c}, z^m, y) - \frac{f_X(x)f_{X^m,Y}(z^m, y)}{f_{X^m}(x^m)}\Big\}\,dz^m dy \\ &= \frac{\mu_{0,m^c}}{h^{d_m}}\int_{\mathbb{R}^d} K_m\big(\frac{z^m - x^m}{h}\big)\Big\{f_{X^{m^c},X^m}(x^{m^c}, z^m) - \frac{f_X(x)f_{X^m}(z^m)}{f_{X^m}(x^m)}\Big\}\,dz^m \\ &= \mu_{0,m^c}\int_{\mathbb{R}^d} K_m(u)\Big\{f_{X^{m^c},X^m}(x^{m^c}, x^m + hu) - \frac{f_X(x)f_{X^m}(x^m + hu)}{f_{X^m}(x^m)}\Big\}\,du \\ &= \mu_{0,m^c}\int_{\mathbb{R}^d} K_m(u)\Big\{f_{X^{m^c},X^m}(x^{m^c}, x^m + hu) - \frac{f_X(x)f_{X^m}(x^m + hu)}{f_{X^m}(x^m)}\Big\}\,du \\ &\le Lh\mu_{0,m^c}\mu_{1,m}\Big\{1 + \frac{f_X(x)}{f_{X^m}(x^m)}\Big\}.\end{aligned}$$

Moreover, we will show below that

$$\mathrm{Var}\Big\{\frac{\mu_{0,m^c}}{h^{d_m}}K_m\big(\frac{X_i^m - x^m}{h}\big)\Big(f_{X^{m^c}|X^m,Y}(x^{m^c}; X_i^m, Y_i) - \frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big)\Big\} = O\Big(\frac{1}{h^{d_m}}\Big), \qquad (46)$$

uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$. It follows that $|\Pi_2| = o_p(1)$, uniformly for $h \in [n^{-\beta}, n^{-\alpha}]$.

It remains to show (44) and (46). We have that

$$\begin{aligned}
&\mathrm{Var}\Big\{Y_i K_m\big(\frac{X_i^m - x^m}{h}\big)\Big(f_{X^{m^c}|X^m,Y}(x^{m^c}; X_i^m, Y_i) - \frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big)\Big\} \\
&\leq \mathbb{E}\Big[\Big\{Y_i K_m\big(\frac{X_i^m - x^m}{h}\big)\Big(f_{X^{m^c}|X^m,Y}(x^{m^c}; X_i^m, Y_i) - \frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big)\Big\}^2\Big] \\
&\leq \sup_{z^m\in\mathbb{R}^{d_m}, y\in\mathcal{D}_Y}\Big[y^2\Big\{f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y) \\
&\qquad\qquad - \frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}^2 f_{X^m}(z^m)\Big]\int_{\mathbb{R}^{d_m}} K_m^2\big(\frac{z^m - x^m}{h}\big)\,dz^m \\
&= \sup_{z^m\in\mathbb{R}^{d_m}, y\in\mathcal{D}_Y}\Big[y^2\Big\{f_{X^{m^c}|X^m,Y}(x^{m^c}; z^m, y) - \frac{f_X(x)\eta(x)}{f_{X^m}(x^m)\eta_m(x^m)}\Big\}^2 f_{X^m}(z^m)\Big]\nu_m h^{d_m}
\end{aligned}$$

Thus, (44) holds since the density functions are bounded. The claim in (46) can be seen by the same argument.

The proof is competed by combining Part I, Part II and Slutsky's Theorem. ■

References

Anderson, T. W. (1957) Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Assoc.,* **52**, 200-203.

Biau, G. and Devroye, L. (2015) *Lectures on the Nearest Neighbour Method.* Springer Series in the Data Sciences, Springer International Publishing Switzerland.

Breiman, L. (2002) Random forests. *Machine Learning*, **45**, 5–32.

Cai, T. T. and Zhang, A. (2016) Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis* **150**, 55–74.

Cai, T. T. and Zhang, L. (2018) High-dimensional linear discriminant analysis: optimality, adaptive algorithms and missing data. *J. Roy. Statist. Soc., Ser. B*, **81**, 675–705.

Carroll, R. J., Ruppert, D. and Walsh, A. H. (1998) Local estimating equations. *J. Amer. Statist. Assoc.*, **93**, 214-227.

Chen, Y.-H. and Chen, H. (2000) A unified approach to regression analysis under double-sampling designs. *J. Roy. Statist. Soc., Ser. B*, **62**, 449–460.

Dempster, A. P., Laird, N. M. and Rudin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., Ser. B (with discussion)*, **39**, 1–38.

Elsener, A. and van de Geer, S. (2018) Sparse spectral estimation with missing and corrupted measurements. *Stat*, **8**, 1–11.

Fan, J. and Gijbels, I. (1996) *Local polynomial modelling and its applications*, Chapman & Hall/CRC, Boca Raton, Florida.

Ford, B. L. (1983) An overview of hot-deck procedures. In Madow, W. G., Olkin, I. and Rubin, D. B. (Eds.) *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies*, 185–207. Academic Press, New York.

Fuller, W. A. (1998) Replication variance for two-phase samples. *Statistica Sinica*, **8**, 1153–1164.

Horvitz, D. G. and Thompson, D. J. (1956) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.

Janson, S. (1984). The asymptotic distributions of incomplete U-statistics, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **66**, 495–505.

Jiang, X., Jiang, J. and Liu, Y. (2011) Nonparametric regression under double-sampling designs. *J. Syst. Sci. Complex*, **24**, 167–175.

Josse, J. and Reiter, J. P. (2018) Introduction to the special section on missing data. *Statistical Science*, **33**, 139–141.

Kang, J. D. Y., and Schafer, J. L. (2007) Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, **22**, 523–539.

Lin, H.-W. and Chen Y.-H. (2014) Adjustment for missing confounders in studies based on observational databases: 2-Stage calibration combining propensity scores from primary and validation data. *American Journal of Epidemiology*, **180**, 308–317.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical analysis with missing data.* Wiley, New Jersey.

Loh, P.-L. and Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. *Ann. Statist.*, **40**, 1637–1664.

Lounici, K. (2014) High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, **20**, 1029-1058.

Miao, W., Ding, P. and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.*, **111**, 1673-1683.

Molenberghs, G., Fitzmaurice, G., Kenwood, M. G., Tsiatis, A. and Verbeke, G. (2015) *Handbook of Missing Data Methodology.* CRC Press, Florida.

Parzen, E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.

Pantanowitz, A. and Marwala, T. (2009) Missing data imputation through the use of the random forest algorithm. *Adv. in Comp. Intel.*, **116**, 53–62.

Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.

Rubin, D. B. (1976) Inference and missing data. *Biometrika*, **63**, 581–592.

Tsiatis, A. (2006) *Semiparametric Theory and Missing Data*. Springer Series in Statistics, Spinger-Verlag New York.

Tsybakov, A. B. (2004) *Introduction to nonparametric estimation.* Springer series in statistics, Springer, New York.

van der Vaart, A. (1998) *Asymptotic Statistics*. Cambridge University Press, Cambridge, U.K.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., Schouten, R., Gaffert, P., Meinfelder, F. and Gray, B. (2018) `mice`: Multivariate imputation via chained equations. `R` package, available from `CRAN`.

Wahba, G. (1990) *Spline Models for Observational Data*. SIAM, Philadelphia, PA.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Chapman and Hall/CRC, Boca Raton, FL.

Wang, S., Shao, J., and Kim, J. (2014) An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, **24**, 1097-1116.

Yang, S. and Ding, P. (2020) Combining multiple observational data sources to estimate causal effects. *J. Amer. Statist. Assoc.*, **115**, 1540-1554.

Zhang, A., Brown, L. D. and Cai, T. T. (2019) Semi-supervised inference: general theory and estimation of means. *Ann. Statist.*, **47**, 2538–2566.

Zhu, Z., Wang, T. and Samworth, R. J. (2019) High-dimensional principal component analysis with heterogeneous missingness. *Preprint*, `ArXiv:1906.12125`.