Simple nonlinear imputation

December 2020

1 Introduction

In this paper we consider GMM estimation in models where an explanatory variable is missing. The researcher observes a random sample of LHS and RHS variables ($Y_i \in \mathbb{R}$) and ($X_i \in \mathbb{R}, Z_i \in 1 \times \mathbb{R}^k$), respectively, and aims to estimate a parameter value $\beta \in B$, where B is a known compact subset of $\mathbb{R}^p$. There is a known, bounded vector-valued function with continuous and bounded derivatives

$$g_0 : Supp(Y_i) \times Supp(Z_i) \times Supp(X_i) \times B \to \mathbb{R}^q,$$

such that its expected value is constant zero conditional on X_i, Z_i if and only if it is evaluated at the true parameter values (β):

$$E[g_0(Y_i, X_i, Z_i; \beta)|X_i, Z_i] = 0. \tag{1}$$

The key problem in our setting is that the scalar X_i variable is missing whenever M_i, the missingness indicator is 1.

Example 1 (Probit Non-linear Least Squares). Assume that the data generating process is the same as in the previous example (equations (2)-(4)). Another way to estimate the coefficients is via non-linear least squares, when the set of population moments are

$$E\left[\begin{array}{c} X_i(Y_i - \Phi(aX_i + Z_ic)) \\ Z_i(Y_i - \Phi(aX_i + Z_ic)) \end{array}\right].$$

Example 2 (Probit Maximum Likelihood). The researcher collected data of a

binary outcome variable Y and the vector of independent variables Z for a large sample, but only has information about a control variable X in a subsample.[1] In the following we denote the cumulative distribution function of the standard normal distribution by Φ and the probability density function by ϕ. The following model connects the observables:

$$Y_i = \mathbf{1}[\alpha X_i + Z_i\gamma > \epsilon_i] \tag{2}$$

$$\epsilon_i \sim N[0,1] \tag{3}$$

$$\epsilon \perp X_i, Z_i \tag{4}$$

One way to estimate the coefficients is based on the likelihood principle, which corresponds to the population moment (g_0)

$$E\left[\begin{array}{c} X_i\left(Y_i\frac{\phi(aX_i+Z_ic)}{\Phi(aX_i+Z_ic)} - (1-Y_i)\frac{\phi(aX_i+Z_ic)}{1-\Phi(aX_i+Z_ic)}\right) \\ Z_i\left(Y_i\frac{\phi(aX_i+Z_ic)}{\Phi(aX_i+Z_ic)} - (1-Y_i)\frac{\phi(aX_i+Z_ic)}{1-\Phi(aX_i+Z_ic)}\right) \end{array}\right].$$

LITERATURE REVIEW, SELLING MISSING HERE

2 A simple imputation GMM estimator

Next we define three GMM estimators:

1. The infeasible *full-data GMM estimator* is based on the population moment
$$E[g_0(Y_i, Z_i, X_i; \beta)] = 0.$$

2. The *complete case GMM estimator* is based on
$$E\left[(1-M_i)g_0(Y_i, X_i, Z_i; \beta)\right] = 0. \tag{5}$$
for $\tilde{g} : Supp(Y_i, Z_i, X_i, M_i) \times B$ for $B \subset \mathbb{R}^p$,

3. For variables $Z_i^1 \subset Z_i$ with support on $\mathbb{R}^{k_1}$, the *imputation GMM estimator* is defined with population moments
$$\left[\begin{array}{c} (1-M_i)g_0(Y_i, X_i, Z_i; \beta)] \\ M_i e(Y_i, Z_i^1; \beta) \end{array}\right] = 0, \tag{6}$$

[1] The method also works if the researcher does not

with $e : \mathbb{R} \times \mathbb{R}^{k_1} \times B \to \mathbb{R}$ given by

$$E[e(y, z^1, b)] = E[E[g_0(Y_i, X_i, Z_i, b)|Y_i = y, Z_i^1 = z^1]]. \tag{7}$$

The infeasible *full-data GMM estimator* has the sample moment

$$\hat{g}_0(b) = n^{-1} \sum_{i=1}^{n} g_0(y_i, x_i, z_i, b) \tag{8}$$

and given a weighting matrix $\hat{W}_0 \xrightarrow{p} W_0$ (positive definite) minimizes

$$\hat{Q}_n^0(b) = \hat{g}_0(b)' \hat{W}_0 \hat{g}_0(b) \tag{9}$$

with respect to b.

The *complete case estimator* is the result of a usual strategy of omitting the observations with missing values. This estimator is defined by the sample moment

$$\tilde{g}(b) = n^{-1} \sum_{i=1}^{n} (1 - m_i) g_0(y_i, x_i, z_i; b), \tag{10}$$

and it is the M-estimator minimizing

$$\tilde{Q}_n(b) = \tilde{g}(b)' \hat{\tilde{W}} \tilde{g}(b) \tag{11}$$

with respect to b, where the $\hat{\tilde{W}}$ is a symmetric weighting matrix such that for some $\tilde{W}$ (positive definite)

$$\hat{\tilde{W}} \xrightarrow{p} \tilde{W}. \tag{12}$$

In addition to the feasible identifying moments ($\tilde{g}$) we also add imputation moments to our *imputation GMM estimator* in order to increase efficiency. We define the function

$$g(y, x, z, m; b, e) = \begin{bmatrix} (1 - m) \cdot g_0(y, x, z; b)] \\ m \cdot e(y, z^1; b) \end{bmatrix}, \tag{13}$$

where the last argument is a function that represents the conditional expectation of $g_0(Y_i, X_i, Z_i, b)$ given $Z_i^1 = z^1$ and $Y_i = y$. Define the sample analogues (for

a sample of size n)

$$\hat{g}(b,\hat{e}) = n^{-1}\sum_{i=1}^{n} g(y_i, x_i, z_i, m_i; b, \hat{e}) = \tag{14}$$
$$= n^{-1}\sum_{i=1}^{n}\begin{bmatrix} (1-m_i)\cdot g_0(y_i, x_i, z_i; b) \\ m_i \cdot \hat{e}(y_i, z_i^1; b)) \end{bmatrix},$$

where $\hat{e}$ is an estimator of the conditional expectation e. In this paper we give the Nadaraya-Watson estimator as a specific example for a viable $\hat{e}$, but other, potentially more sophisticated estimators could work as well. The *imputation GMM estimator* is minimizing

$$\hat{Q}_n(b) = \hat{g}(b;\hat{e})'\hat{W}\hat{g}(b;\hat{e}) \tag{15}$$

with respect to b, where the $\hat{W}$ is a symmetric weighting matrix such that for some W (positive definite)

$$\hat{W} \xrightarrow{p} W. \tag{16}$$

2.1 Moments additive in X_i and Y_i

In order to preserve the simplicity of the approach of Abrevaya and Donald (2017), the applied researcher may only want to focus on the moments that are additive in X_i and Y_i. If a particular row of g_0 can be written as

$$g_0(Y_i, X_i, Z_i, b) = h_1(Y_i, Z_i) + h_2(X_i, Z_i)$$

for some known h_1, h_2 functions, then

$$E[M_i E[g_0(Y_i, X_i, Z_i, b)|Y_i, Z_i^1]] = E\left[M_i(h_1(Y_i, Z_i) + E[h_2(X_i, Z_i)|Z_i^1])\right], \tag{17}$$

due to[2] the Law of Iterated Expectations and because conditional on Z_i^1, we have that missingness is independent of X_i, Z_i^0 and Y_i. If we use additive moments, we only need to estimate the conditional expectation of the component that contains the missing variable X_i conditional on Z_i^1, excluding Y_i - thereby decreasing the noise we introduce due to imputation. The length of the conditioning vector is a key factor that determines the performance of imputation in

[2] A more detailed calculation is available in the Appendix.

section 2.2.

Example 1: Probit with NLS (continued)

The imputation moments we add are

$$E\begin{bmatrix} M_i E[X_i|Z_i^1, Y_i](Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1, Y_i]) \\ M_i E[Z_i^0|Z_i^1, Y_i](Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1, Y_i]) \\ M_i Z_i^1(Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1, Y_i]) \end{bmatrix}. \tag{18}$$

As argued above, for the additive elements of this vector we may use the identical moments

$$E\begin{bmatrix} M_i E[X_i|Z_i^1, Y_i](Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1]) \\ M_i Z_i^0(Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1]) \\ M_i Z_i^1(Y_i - E[\Phi(aX_i + Z_i c)|Z_i^1]) \end{bmatrix}, \tag{19}$$

which are more simple to estimate. Following Abrevaya and Donald (2017), we only use the additive moments (last two elements of the matrix above) in the Monte Carlo simulations. We get the parametrized case of the authors if the conditional expectation is further parametrized by a finite dimensional parameter vector, which also ensures it can be estimated $\sqrt{n}$-consistently under regulatory conditions.

2.2 Comparison of asymptotic properties

In the following arguments we closely follow Ichimura and Newey (2015) and Chernozhukov et al. (2018). We denote the conditional pdf of a variable X_i given and another variable Z_i as $f_{x|z}$. Moreover, we define notation for the Jacobians of the population moments of the various estimators along with the

corresponding sample analogues.

$$
\begin{aligned}
G_0 &= \left.\frac{\partial E[g_0(Y_i, Z_i, X_i; b)}{\partial b}\right|_{b=\beta} \\
\hat{G}_0 &= \left.\frac{\partial \hat{g}_0(b)}{\partial b}\right|_{b=\beta}, \\
\tilde{G} &= \left.\frac{\partial E[(1 - M_i) \cdot g_0(Y_i, Z_i, X_i; b)}{\partial b}\right|_{b=\beta} \\
\hat{\tilde{G}} &= \left.\frac{\partial \tilde{g}(b)}{\partial b}\right|_{b=\beta} \\
G &= \left.\frac{\partial E[g(Y_i, X_i, Z_i, M_i; b, e)}{\partial b}\right|_{b=\beta} \\
\hat{G} &= \left.\frac{\partial \hat{g}(b, \hat{e})}{\partial b}\right|_{b=\beta}
\end{aligned}
$$

We denote the imputation estimator as a random variable by β_n and the complete case estimator by $\tilde{\beta}_n$. We further define V_i as a subset of the joint vector (Y_i, Z_i^1) which we condition on to calculate an infinite dimensional nuisance parameter for the imputation GMM in the first stage. In the generic case $V_i = (Y_i, Z_i^1)$, while in the additive moments case $V_i = Z_i^1$.

We use two sets of assumptions. The baseline assumptions are necessary to derive standard results for the complete case estimator.

Assumption 1.

a) $E[g_0(Y_i, X_i, Z_i, b)|X_i, Z_i, M_i = 0] = 0 \ (X_i, Z_i) - a.s. \Leftrightarrow b = \beta.$

b) $G_0' W_0 G_0$ *is invertible.*

c) $P[M_i = 1|X_i, Z_i] < 1 \ X_i, Z_i - a.s.$

The first two conditions in Assumption 1 ensure identification in the case with missingness. Condition 1/a) is satisfied if the missingness is not dependent on Y_i, conditional on X_i, Z_i. We call this assumption the missing-at-random (MAR) assumption. Part b) is a usual condition for the infeasible full-data GMM estimator expressing that none of the moments are redundant, and in our setting it is equivalent to require that G_0 is full rank (see for example Newey and McFadden 1996).

We contrast the baseline Assumption 1 with the following set of conditions:

Assumption 2.

a) In addition to condition 1/a), there is a partitioning of $Z_i = (Z_i^0, Z_i^1)$ *with* $Supp(Z_i^1) \subset \mathbb{R}^{k_1}$ *such that*

$$E[g_0(Y_i, X_i, Z_i, \beta)|Y_i, Z_i^1, M_i = 0] = E[g_0(Y_i, X_i, Z_i, \beta)|Y_i, Z_i^1].$$

b) $G'WG$ *is invertible.*

c) $P[M_i = 1|X_i, Z_i] < 1$ $X_i, Z_i - a.s.$

Condition 2/a)-b) are strengthened versions of the analogue conditions in Assumption 1. The typical sufficient condition for 2/a) is a strengthened version of the MAR assumption, that the missingness is independent of Y_i and X_i, Z_i^0, conditional on Z_i^1. However, this is the weakened version of the missing-at-completely-random assumption often assumed by researchers. Condition 2/b) once again ensures that the added imputation moments represent new information on the limit, as this assumption is equivalent to requiring that G is full rank. Condition 2/b) rules out the case when Z_i^1 and Y_i is independent of X_i, which means that they do not provide any information about X_i.

In addition, we make regulatory assumptions on the estimator $\hat{e}$ which make the asymptotic theory calculations feasible.

Assumption 3.

a) Almost surely, $\hat{e}$ *is a* $k_v + 1$*-times continuously differentiable function of* (Y_i, Z_i^1) *with the* k_v+1*-times derivative bounded with probability approaching to 1 as* $n \to \infty$*.*

b) The expectation of $M_i\hat{e}(Y_i, Z_i^1, \beta)$ *over* $\hat{e}$ *has a smooth linear representation in the non-missing observational values. There is a sequence of* $k_v + 1$*-times differentiable* $h_n : Supp((Y_i, X_i, Z_i)) \to \mathbb{R}^q$ *deterministic functions with the* $k_v + 1$*-times derivative bounded as* $n \to \infty$ *such that*

$$\int \hat{e}(y, z^1, \beta) f_{y,z^1|m=1}(y, z^1) d(y, z^1) = \frac{1}{n} \sum_j (1 - M_j) h_n(Y_j, X_j, Z_j) + o_p(\sqrt{n}).$$

Condition 3/a) ensures that the possible set of estimating functions is not too complex. It serves the purpose of a usual stochastic equicontinuity condition

and it can be replaced by other conditions of the same kind (see in Van der Waart and Wellner 1996 or Andrews 1994a). Condition 3/b) restricts the type of estimators we consider to estimators that behave close to linear, asymptotically. We show that the Nadaraya-Watson estimator satisfies this condition. This condition is Assumption 3.8 in Newey (1994), but similar conditions are also explored in Ichimura and Newey (2015) or Cattenao and Jansson (2018). Assumption 3 can be weakened or substituted with alternative set of assumptions that would accept a different set of estimators (for example the kernel estimator with uniform kernel). However, our focus is on the comparison of Assumption 1 and 2, which are assumptions on the data generating process instead of the chosen estimation procedure for the nuisance parameter $\hat{e}$.

Proposition 1. *Under Assumption 1, the complete case GMM estimator $\tilde{\beta}_n$ is consistent and*

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N[0, (\tilde{G}'W\tilde{G})^{-1}\tilde{G}'W\tilde{\Omega}W\tilde{G}(\tilde{G}'W\tilde{G})^{-1}],$$

where $\tilde{\Omega} = \lim_n E\left[n^{-1}\sum_i (1-m_i)g_0(y_i, x_i, z_i; \tilde{\beta}_n)' g_0(y_i, x_i, z_i; \tilde{\beta}_n)\right]$.

Under Assumption 2-3, given that $E[e(y,z,\beta) - \hat{e}(y,z,\beta)] = o_p(n^{-1/2})$, the imputation estimator β_n is consistent and

$$\sqrt{n}(\beta_n - \beta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = \lim_n E\left[n^{-1}\sum_i g(y_i, x_i, z_i, m_i; b, \hat{e})' g(y_i, x_i, z_i, m_i; b, \hat{e})\right]$.

The proposition states that the imputation estimator has a variance that approaches zero with the parametric rate, but the bias term form the first stage may prevent $\sqrt{n}$-convergence. Typically, the bias of the first stage can be driven to zero with a higher rate than $\sqrt{n}$ with undersmoothing. We also highlight an estimation method for the variance-covariance matrix by specifying the Ω this way.

2.2.1 First stage using the Nadaraya-Watson estimator

We estimate the conditional expectation by a Nadaraya-Watson type estimator. For simplicity of exposition, we assume that all observables are continuously

distributed, and we also denote $V_i = (Y_i, Z_i^1)$, the joint vector in $\mathbb{R}^{k_v}$. We define

$$\hat{e}(y_i, z_i^1; b) = \frac{\sum_{j:M_j=0} K[H^{-1}(v_i - v_j)] g_0(y_j, x_j, z_j, b)}{\sum_j K[H^{-1}(v_i - v_j)]}. \tag{20}$$

For the sake of simplicity, we will assume that H is a diagonal matrix with positive diagonal entries. Let us have the entry that decreases to zero at the slowest rate denoted by h_{max}, moreover let us write $\prod h_k = h$.

Remark 1. When any of the variables are discrete, there is no need for kernel smoothing, and we would need to calculate the averages for each value of the discrete variable separately. In that case the discrete variable does not contribute to the curse of dimensionality, so k_v should be understood as the number of continuous random variables in V_i in our statements below.

Assumption 4.

a) *K is a $k_v + 1$-times continuously differentiable symmetric kernel with*

$$K(u)||u||_\infty^{k_v+1+\delta} \to 0$$

for some $\delta > 0$.

b) *$h_{max} = o\left(n^{-\frac{2+\delta}{6+4\delta}}\right)$, but $nh \to \infty$.*

c) *The joint pdf of V_i conditional on $M_i = 1$ is continuous and bounded away from zero and infinity on the support.*

d) *The conditional distribution function $f_{v_j|v_i}(v_j)$ for a V_j chosen from the fully observed subsample and V_i chosen from the subsample with missing values, is twice differentiable with bounded Hessian for every conditioning value $V_i = v_i$.*

Some of these assumptions are stronger than necessary (notably, conditions 4/c) and 4/d), but they simplify the algebra greatly in the usual bias calculations of the non-parametric estimator. Condition 4/a-b) are needed to establish the stochastic equicontinuity conditions and uniform consistency. In our simulations we use a kernel with an exponential tail, so conditions a) and b) are satisfied with the usual $h_{max} = o(n^{-1/4})$ and $nh \to \infty$ bounds.

Corollary 1. *If $\hat{e}$ is the Nadaraya-Watson estimator with $k_v < 4$, under As-*

sumption 2, 4

$$\sqrt{n}(\beta_n - \beta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = \lim_n E[n\hat{g}(\beta_n, \hat{e})'\hat{g}(\beta_n, \hat{e})]$.

The corollary suggests that we need $k_v < 4$, if every term in the V_i vector is continuous (including the LHS variable). There are two reasons to include an (always observed) RHS variable into the imputation moments as conditioning variable:

1. Weakening of the missing-at-completely-random assumption: we think the variable is related to missingness.
2. Predictive power for X_i: observing the variable gives information about the missing RHS variable.

Even if the second point would not warrant an inclusion of a particular element of Z_i into the group of conditioning variables in $\hat{e}$, if we think that a factor may be related to missingness, it needs to be included in the first stage.

Example 2: Probit with Maximum Likelihood (continued)

We add the population moments

$$E\left[\begin{array}{c|c} M_i X_i \left(Y_i \frac{\phi(aX_i+Z_ic)}{\Phi(aX_i+Z_ic)} - (1-Y_i)\frac{\phi(aX_i+Z_ic)}{1-\Phi(aX_i+Z_ic)}\right) & Z_i^1 \\ M_i Z_i \left(Y_i \frac{\phi(aX_i+Z_ic)}{\Phi(aX_i+Z_ic)} - (1-Y_i)\frac{\phi(aX_i+Z_ic)}{1-\Phi(aX_i+Z_ic)}\right) & Y_i \end{array}\right].$$

Here we need to condition on Z_i^1 and Y_i as well, when calculating the conditional expectation of the moments. However, Y_i is not going to require kernel smoothing, since it is discrete.

2.3 The role of the weighting matrix and efficiency

Our goal is to minimize the Mean Squared-Error (MSE), which can be calculated as the expected value of the diagonal of the matrix

$$\begin{aligned}(\beta_n - \beta)(\beta_n - \beta)' &= (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{g}(\beta, \hat{e})((\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{g}(\beta, \hat{e}))' = \\ &= (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}_0\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1},\end{aligned} \quad (21)$$

for

$$\hat{\Omega}_0 = \hat{g}(\beta, \hat{e})\hat{g}(\beta, \hat{e})'.$$

If we set

$$\hat{W}^{-1} = \hat{\Omega}_0,$$

then

$$(\beta_n - \beta)(\beta_n - \beta)' = \left(\hat{G}'\hat{\Omega}_0^{-1}\hat{G}\right)^{-1}. \tag{22}$$

We note that by the usual arguments in the GMM literature,

$$diag\left((\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}_0\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1} - (\hat{G}'\hat{\Omega}_0^{-1}\hat{G})^{-1}\right) \geq 0 \; ev., \tag{23}$$

which means this is the infeasible optimal weighting. This optimal weighting matrix can be estimated by

$$\hat{\Omega} = n^{-1}\sum g(Y_i, X_i, Z_i, M_i, \beta_n)g(Y_i, X_i, Z_i, M_i, \beta_n)',$$

which is a block-diagonal matrix that can be written as

$$\hat{\Omega} = \begin{bmatrix} \hat{\tilde{\Omega}} & 0 \\ 0 & \hat{B} \end{bmatrix}. \tag{24}$$

The block matrix $\hat{\tilde{\Omega}}$ is the estimate for the inverse of the optimal weighting matrix for the complete case GMM estimator. The block matrix corresponding to the imputation moments is $\hat{B}$. This matrix is positive definite if the additional moments do not have a zero optimal weight as n tends to infinity, which is guaranteed by Assumptions 2-3. In this case $diag(G\hat{W}G)^{-1}$ is smaller or equal than the diagonal of the optimal MSE of the estimator that does not contain the added moments, (which is $diag(\tilde{G}'\hat{\tilde{\Omega}}^{-1}\tilde{G})^{-1}$).

Proposition 2. *Under Assumption 2-3* $MSE(\beta_n) \leq MSE(\tilde{\beta}_n)$ *ev. for*[3] *any admissible weighting of the* $\tilde{\beta}_n$ *estimator, with the inequality being strict for at least one element of* β*.*

The proposition states that in large samples the imputation estimator will always increase efficiency, if the optimal weighting matrix calculated as prescribed

[3] An event is true eventually means that there is an $N < \infty$ such that the event is true for every $n > N$. We abbreviate this as "ev."

above does not exclude the additional imputation moments as $n \to \infty$. This can happen only if the Z_i^1 has too high dimensions (so we do not have $\sqrt{n}$-consistency in general) or if the X_i is independent of the observables that are always observed.

Remark 2. Under our assumptions, if an element of $\hat{g}$ does not converge with a $\sqrt{n}$ rate to zero due to the estimated nuisance parameter, its relative weight is set to be arbitrarily close to zero by the optimal weighting matrix, eventually. However, if the applied researcher implements the imputation method for too many dimensions, imputation may slightly *increase* standard errors in finite samples.

Remark 3. Our estimation method and theory can be easily extended to the case when there are more than one RHS variables are missing. However, we do not pursue the full imputation estimator where the researcher uses two variables with missing values to predict each other ('Swiss cheese case').

3 Monte Carlo simulation

4 Conclusion

5 References

Abrevaya and Donald (2017) Ichimura and Newey (2015) Chernozhukov et al. (2018) Cattenao amd Jansson (2018)

6 Appendix

6.1 Additive g_0

Here we show how one can achieve dimension reduction when the moments are additive in Y_i and X_i.

$$\begin{aligned} & E[M_i E[g_0(Y_i, X_i, Z_i, b)|Y_i, Z_i^1]] \\ & = E[M_i E[h_1(Y_i, Z_i) + h_2(X_i, Z_i)|Y_i, Z_i^1]] \\ & = E[M_i h_1(Y_i, Z_i)] + E[M_i E[h_2(X_i, Z_i)|Y_i, Z_i^1]] \\ & = E[M_i h_1(Y_i, Z_i)] + E[E[M_i E[h_2(X_i, Z_i)|Y_i, Z_i^1]|Z_i^1]] \\ & = E[M_i h_1(Y_i, Z_i)] + E[E[M_i|Z_i^1]E[h_2(X_i, Z_i)|Z_i^1]] \\ & = E[M_i h_1(Y_i, Z_i)] + E[M_i E[h_2(X_i, Z_i)|Z_i^1]] \\ & = E\left[M_i(h_1(Y_i, Z_i) + E[h_2(X_i, Z_i)|Z_i^1])\right] \end{aligned} \tag{25}$$

6.2 Proposition 1

Under Assumption 1, the complete case GMM estimator $\tilde{\beta}_n$ is consistent and

$$\sqrt{n}(\tilde{\beta}_n - \beta) \xrightarrow{d} N[0, (\tilde{G}'W\tilde{G})^{-1}\tilde{G}'\tilde{W}\tilde{\Omega}\tilde{W}\tilde{G}(\tilde{G}'\tilde{W}\tilde{G})^{-1}],$$

where $\tilde{\Omega} = \lim_n E[n\tilde{g}(\beta)'\tilde{g}(\beta)]$.
Under Assumption 2-3, given that $E|e(y,z,b) - \hat{e}(y,z,b)| = o_p(n^{-1/2})$ for all $b \in B$, the imputation estimator β_n is consistent and

$$\sqrt{n}(\beta_n - \beta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = \lim_n E[n\hat{g}(\beta_n, \hat{e})'\hat{g}(\beta_n, \hat{e})]$.

Writing up the first order Taylor expansion of $\hat{g}$ around β gives

$$\begin{aligned} 0 = \hat{G}'(\beta_n; \hat{e})\hat{W}\hat{g}(\beta_n; \hat{e}) = \\ = \hat{G}'(\beta_n; \hat{e})\hat{W}\hat{g}(\beta; \hat{e}) + \hat{G}'(\beta_n; \hat{e})\hat{W}\hat{G}(\bar{\beta}_n; \hat{e})(\beta_n - \beta) = \\ = \hat{G}'\hat{W}\hat{g}(\beta; \hat{e}) + \hat{G}'\hat{W}\bar{G}(\beta_n - \beta). \end{aligned} \tag{26}$$

We abbreviated the notation for the various matrices from the second row. Here $\bar{\beta}_n$ is a vector of convex combinations of β and $\hat{\beta}_n$, but this value can (and generally should) be different for different rows of $\hat{G}$. In this sense we abuse the

notation for $\hat{G}$ somewhat when we say it is evaluated at $\bar{\beta}$.

Given the boundedness of g_0 and that it is continuously differentiable Assumption 2, $||\hat{G}'\hat{W}\bar{G}||$ is bounded and $\hat{g}(\beta;\hat{e}) \to 0$ with probability approaching 1. This in turn yields

$$\begin{aligned}\beta_n - \beta &= -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\hat{g}(\beta;\hat{e}) \\ &= -(G'WG)^{-1}G'W\hat{g}(\beta;\hat{e}) + o_p(\hat{g}(\beta;\hat{e})).\end{aligned} \tag{27}$$

We consider the decomposition

$$\begin{aligned}\sqrt{n}\hat{g}(\beta;\hat{e}) = &\sqrt{n}\left(\hat{g}(\beta;e) - E[g(\beta;e)]\right) \\ &+ \sqrt{n}(\hat{g}(\beta;\hat{e}) - \hat{g}(\beta,e) - E_{Y_i,Z_i^1}[\hat{g}(\beta;\hat{e}) - \hat{g}(\beta,e)]) \\ &+ \sqrt{n}\left(E_{Y_i,Z_i^1}[\hat{g}(\beta,\hat{e})] - E[\hat{g}(\beta,\hat{e})]\right) \\ &+ \sqrt{n}E[\hat{g}(\beta,\hat{e})],\end{aligned} \tag{28}$$

First, we prove that the second term

$$\sqrt{n}(\hat{g}(\beta;\hat{e}) - \hat{g}(\beta,e) - E_{Y_i,Z_i^1}[\hat{g}(\beta;\hat{e}) - \hat{g}(\beta,e)]) = o_p(1). \tag{29}$$

For this, consider that $\hat{e}$ are uniformly bounded, p times continuously differentiable function on a compact domain admitting a valid Taylor-expansion of order $k_{z_1} + 1$ with bounded relictum term even over possible $\hat{e}$ realizations as per Assumption 3 and under $\sqrt{n}h_{max}^2 = o(1)$ as n grows with probability 1. This means that the possible $\hat{e}$ functions is class III set after Andrews (1994c), and so by Theorem 5 of that paper stochastic equicontinuity holds, which gives (29) by definition (see for example page 2251 of the same work) and because $\hat{e}$ converges uniformly to e for every β.

Second we examine the third term,

$$\begin{aligned}&\sqrt{n}\left(E_{Y_i,Z_i^1}[\hat{g}(\beta,\hat{e})] - E[\hat{g}(\beta,\hat{e})]\right) \\ &= \sqrt{n}\left(\sum_i (1-M_i)h_n(Y_i,X_i,Z_i) - E[(1-M_i)h_n(Y_i,X_i,Z_i)]\right) + o_p(1),\end{aligned} \tag{30}$$

due to the representation in Assumption 3/b. Once again, the function class $\mathcal{H} = \{h_n(.) : n \in \mathbb{N}\}$ is of class III as it is defined in Andrews (1994c), since

h_n is itself p-times continuously differentiable and uniformly bounded with a compact domain. It is also totally bounded using the uniform entropy metric by assumption. This together implies that $\mathcal{H}$ is Donsker, which then together with

$$E_{Y_i,Z_i^1}[e(Y_i, Z_i^1, \beta)] = 0 \tag{31}$$

and

$$P[\sup|\hat{e} - e| > \epsilon] \to 0, \tag{32}$$

we get that

$$\sqrt{n}\left(E_{Y_i,Z_i^1}[\hat{g}(\beta, \hat{e})] - E[\hat{g}(\beta, \hat{e})]\right) = o_p(1) \tag{33}$$

by the same argument as above.

As for the first term, consider the imputation moments

$$\begin{aligned} &E[M_i E[g_0(Y_i, X_i, Z_i; \beta)|Z_i^1, Y_i, M_i = 0]] \\ &\quad = E[M_i E[g_0(Y_i, X_i, Z_i; \beta)|Z_i^1, Y_i, M_i = 1]] \\ &\quad = E[P[M_i = 1|Z_i^1]E[g_0(Y_i, X_i, Z_i; \beta)|Z_i^1, M_i = 1]] \\ &\quad = E[P[M_i = 1|Z_i^1]E[E[g_0(Y_i, X_i, Z_i; \beta)|X_i, Z_i, M_i = 1]|Z_i^1, M_i = 1]] \\ &\quad = E[P[M_i = 1|Z_i^1]E[E[g_0(Y_i, X_i, Z_i; \beta)|X_i, Z_i]|Z_i^1, M_i = 1]] = 0 \end{aligned} \tag{34}$$

under the same identifying assumption for which the complete case estimator is consistent and the additional exogeneity assumption in Assumption 2/a). Since M_i, Y_i, X_i, Z_i are i.i.d. draws unconditionally and g is bounded,

$$\sqrt{n}t\left(\hat{g}(\beta; e) - E[g(\beta; e)]\right) \xrightarrow{d} N[0, \sigma_t^2] \tag{35}$$

by a Ljapunov CLT for any t row vector, which means

$$\sqrt{n}\left(\hat{g}(\beta; e) - E[g(\beta; e)]\right) \xrightarrow{d} N[0, V] \tag{36}$$

for some variance-covariance matrix by applying the Cramer-Wold device.

Finally, the fourth term is the bias term, which is assumed to be $o_p\left(\sqrt{n}^{-1}\right)$.

This gives that under our assumptions,

$$\sqrt{n}\hat{g}(\beta, \hat{e}) \xrightarrow{d} N[0, V]. \tag{37}$$

Clearly,

$$V[n\hat{g}(\hat{\beta}, \hat{e})'\hat{g}(\hat{\beta}, \hat{e})] \to V. \tag{38}$$

by the boundedness, continuity assumptions on g and consistency using the DCT.

This concludes the proof for the asymptotic behavior of the imputation estimator. Once we note that the complete case estimator is a special case with $\hat{e}_n(Y_i, Z_i^1, \beta) = 0$ and e being the constant 0 function as well, the same applies for the complete case estimator. □

6.3 Corollary 1

If $\hat{e}$ is the Nadaraya-Watson estimator with $k_v < 4$, under Assumption 2, 4

$$\sqrt{n}(\beta_n - \beta) \xrightarrow{d} N[0, (G'WG)^{-1}G'W\Omega WG(G'WG)^{-1}],$$

with $\Omega = \lim_n E[n\hat{g}(\beta_n, \hat{e})'\hat{g}(\beta_n, \hat{e})]$.

This is a direct corollary of Proposition 1 after checking the assumptions that the bias term can be $o(\sqrt{n}^{-1})$ and Assumption 3.

Uniform smoothness condition on $\hat{e}$

Rate of the bias

First we check the rate of bias. The arguments for this part are similar to the usual bias calculations for the Nadaraya-Watson estimator's bias, although we do not have absolute values here.

$$\begin{aligned} &\hat{e}(Y_i, Z_i^1; b) - e(Y_i, Z_i^1; b) \\ &= \frac{(nh)^{-1}\sum_j K[H^{-1}(V_i - V_j)]\left(g_0(Y_j, X_j, Z_j, b) - E[g_0(Y_i, X_i, Z_i, b)|V_i]\right)}{(nh)^{-1}\sum_j K[H^{-1}(V_i - V_j)]}, \end{aligned} \tag{39}$$

where the denominator clearly converges in probability to $f_v(V_i)$, uniformly under Assumption 4, so we focus on the expected value of the numerator.

Let us calculate

$$E\left[E\left[\sum_j K[H^{-1}(V_i - V_j)]\left(g_0(Y_j, X_j, Z_j, b) - E[g_0(Y_i, X_i, Z_i, b)|V_i]\right)|\mathbf{V}\right]\right]$$
$$= E\left[\sum_j K[H^{-1}(V_i - V_j)]\left(E[g_0(Y_j, X_j, Z_j, b) - g_0(Y_i, X_i, Z_i, b)|V_i, V_j]\right)\right], \tag{40}$$

by the assumption that the $Y_i, X_i Z_i$ vectors are i.i.d. This is achieved by taking the second order Taylor-expansion. We also substitute later

$$H^{-1}(V_i - v_j) = \Delta v.$$

$$\sum_j K[H^{-1}(V_i - V_j)]\left(E[g_0(Y_j, X_j, Z_j, b) - g_0(Y_i, X_i, Z_i, b)|V_i, V_j]\right) \tag{41}$$
$$= \sum_j K[H^{-1}(V_i - V_j)]\left(E[D_V g_0(Y_i, X_i, Z_i, b)(V_j - V_i)\right.$$
$$\left.+(V_j - V_i)' D_V^2 g_0(Y_i, X_i, Z_i, b)(V_j - V_i)|V_i, V_j]\right)$$

For the first term,

$$E\left[\sum_j K[H^{-1}(V_i - V_j)]E[D_V g_0(Y_i, X_i, Z_i, b)(V_j - V_i)|V_i, V_j]|V_i\right] \tag{42}$$
$$= nD_V g_0(Y_i, X_i, Z_i, b)\int_{Supp(V_j)} K[H^{-1}(V_i - v_j)](v_j - V_i) f_{v_j|v_i}(v_j)dv_j$$
$$= hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v f_{v_j|v_i}(V_i - H\Delta v)d\Delta v$$
$$= hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v f_{v_j|v_i}(V_i)d\Delta v$$
$$- hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v D f_{v_j|v_i}(v^*)H\Delta v d\Delta v.$$

Since the conditional pdf $f_{v_j|v_i}$ is continuous (Assumption 4) on a compact support (Assumption 1 in addition), it will attain its minimum and maximum. Since $D_V g_0(Y_i, X_i, Z_i, b)$ also has a finite essential supremum, this gives that

the second term in (42) is $O_p(h_{max}^2)$. Moreover, since the integral only depends on V_i in determining $f_{v_i|v_j}$ which is again continuous in the v_i as well (as in addition f_{v_i} is bounded away from zero), we get that

$$E\left[hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v D f_{v_j|v_i}(v^*)H\Delta v d\Delta v\right]$$
$$= hn \cdot O_p(h_{max}^2) \tag{43}$$

unconditionally too.

As for the first term in (42) (keep in mind, we conditioned on V_i at that point yet),

$$\begin{aligned} &hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v f_{v_j|v_i}(V_i)d\Delta v \qquad (44)\\ &= hnD_V g_0(Y_i, X_i, Z_i, b) f_{v_j|v_i}(V_i)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v d\Delta v\\ &= nh \cdot O_p(h_{max}) \end{aligned}$$

for every value V_i can take, as g_0 has bounded derivatives, and $\int K(x)x d\mu(x) < \infty$ for any measure μ. However, if $B_{\epsilon_m}(V_i) \in Supp(V_i)$, we can use that K is symmetric to get

$$\begin{aligned} &hnD_V g_0(Y_i, X_i, Z_i, b)\int_{H^{-1}(Supp(V_j)-V_i)} K[\Delta v]H\Delta v f_{v_j|v_i}(V_i)d\Delta v \qquad (45)\\ &= hnD_V g_0(Y_i, X_i, Z_i, b) f_{v_j|v_i}(V_i)\int_{H^{-1}(Supp(V_j)-V_i)\backslash B_{\epsilon_m}(0)} K[\Delta v]H\Delta v d\Delta v, \end{aligned}$$

for which

$$\left|\int_{H^{-1}(Supp(V_j)-V_i)\backslash B_{\epsilon_m}(0)} K[\Delta v]H\Delta v d\Delta v\right| \leq MK[H^{-1}\mathbf{1}\epsilon_m]\epsilon_m \qquad (46)$$

for some M, since $H^{-1}(Supp(V_j) - V_i)\backslash B_{\epsilon_m}(0)$ is compact, and K is smooth. However, this means that since $K(u)||u||_\infty^{2+\delta} \to 0$, and because $||H^{-1}\mathbf{1}\epsilon_m||_\infty = h_{max}\epsilon_m$,

$$\left|\int_{H^{-1}(Supp(V_j)-V_i)\backslash B_{\epsilon_m}(0)} K[\Delta v]H\Delta v d\Delta v\right| = o(\epsilon_m^{-1-\delta}h_{max}^{2+\delta}). \qquad (47)$$

Now by the law of iterated expectations, the unconditional expectation for the first term in

$$
\begin{aligned}
&E\left[\sum_j K[H^{-1}(V_i - V_j)] D_V g_0(Y_i, X_i, Z_i, b)(V_j - V_i)\right] \qquad (48)\\
&= P[B_{\epsilon_n}(V_i) \in Supp(V_i)] O_p(h_{max}) + P[B_{\epsilon_n}(V_i) \notin Supp(V_i)] o_p(\epsilon_n^{-1-\delta} h_{max}^{2+\delta})\\
&= nh \cdot O(\epsilon_n h_{max}) + nh \cdot o(\epsilon_n^{-1-\delta} h_{max}^{2+\delta}),
\end{aligned}
$$

where the last equality follows from the assumption that $Supp(V_i)$ is compact and the pdf is bounded away from infinity.

With this we arrive to

$$
\begin{aligned}
&E\left[\sum_j K[H^{-1}(V_i - V_j)] E[D_V g_0(Y_i, X_i, Z_i, b)(V_j - V_i)|V_i, V_j]\right] \qquad (49)\\
&\quad = hn \cdot O(h_{max}^2) + nh \cdot O(\epsilon_n h_{max}) + nh \cdot o(\epsilon_n^{-1-\delta} h_{max}^{2+\delta}).
\end{aligned}
$$

We are left with checking the rate of

$$
E\left[K[H^{-1}(V_i - V_j)](V_j - V_i)' D_V^2 E[g_0(Y_i, X_i, Z_i, b)|V_i = v_i^*](V_j - V_i)\right]. \qquad (50)
$$

Once again,

$$
\begin{aligned}
&E\left[\sum_j K[H^{-1}(V_i - V_j)](V_j - V_i)' D_V^2 E[g_0(Y_i, X_i, Z_i, b)|V_i](V_j - V_i)|V_i\right] \qquad (51)\\
&= nh \int K[\Delta v] \Delta v' H' D_V^2 E[g_0(Y_i, X_i, Z_i, b)|V_i = v_i^*] H \Delta v f_{v_j|v_i}(V_i - H\Delta v) d\Delta v,
\end{aligned}
$$

so when we divide by h_{max}^2 and consider that the Hessian of g_0 and the conditional pdf $f_{v_j|v_i}$ is bounded uniformly across V_i by assumption,

$$
\begin{aligned}
&h^{-2}\left|E\left[\sum_j K[H^{-1}(V_i - V_j)](V_j - V_i)' D_V^2 E[g_0(Y_i, X_i, Z_i, b)|V_i](V_j - V_i)|V_i\right]\right|\\
&\leq nhM_2 \int |K[\Delta v] \Delta v' \Delta v| \, d\Delta v\\
&= nhO(1), \qquad (52)
\end{aligned}
$$

for some M_2 constant. This gives that

$$E\left[K[H^{-1}(V_i - V_j)](V_j - V_i)' D_V^2 E[g_0(Y_i, X_i, Z_i, b)|V_i = v_i^*](V_j - V_i)\right] \quad (53)$$
$$= nhO(h_{max}^2).$$

Altogether, this means

$$E\left[(nh)^{-1}\sum_j K[H^{-1}(V_i - V_j)]\left(g_0(Y_j, X_j, Z_j, b) - E[g_0(Y_i, X_i, Z_i, b)|V_i]\right)\right]$$
$$= O(h_{max}^2) + O(\epsilon_n h_{max}) + o(\epsilon_n^{-1-\delta} h_{max}^{2+\delta}). \quad (54)$$

As mentioned above, the denominator in (39) is converging to the pdf of V_i, and as such it is bounded away from zero, so the leading term of $E[\hat{e} - e]$ has the same order as the numerator, which gives

$$E[\hat{e}(Y_i, Z_i^1; b) - e(Y_i, Z_i^1; b)] = O(h_{max}^2) + o(n^{-1/2}) \quad (55)$$

if there are $h_{max}, \epsilon_n \to 0$ pair of sequences such that

$$\epsilon_n h_{max} = o(n^{-1/2}), \quad (56)$$
$$\epsilon_n^{-1-\delta} h_{max}^{2+\delta} = o(n^{-1/2}), \quad (57)$$
$$nh \to \infty. \quad (58)$$

Or to put it in another way, given $h = O(n^{-\eta_h}), \epsilon_n = n^{-\eta_\epsilon}$, for every $\delta > 0$ there are positive η_h, η_ϵ such that

$$\eta_\epsilon + \eta_h > 0.5, \quad (59)$$
$$(2+\delta)\eta_h - (1+\delta)\eta_\epsilon > 0.5, \quad (60)$$
$$\eta_h < 1/k_v. \quad (61)$$

The first two inequalities are valid if

$$0.5 - \eta_h < \eta_\epsilon < \frac{(2+\delta)\eta_h - 0.5}{1+\delta}. \quad (62)$$

Given η_h, we can always choose a suitable η_ϵ iff

$$0.5 - \eta_h < \frac{(2+\delta)\eta_h - 0.5}{1+\delta} \tag{63}$$

$$\iff \frac{1+0.5\delta}{3+2\delta} < \eta_h. \tag{64}$$

Since if $k_v < 4$,

$$\frac{1+0.5\delta}{3+2\delta} < 1/k_v \Leftarrow \delta > 0,$$

this is indeed always possible. However, for $\delta = 2$, we need that $h_{\max} = o(n^{-2/7})$. Note that if we use a kernel with exponential ('light') tails, we need

$$\lim_{\delta\to\infty} \frac{1+0.5\delta}{3+2\delta} = \lim_{\delta\to\infty} \frac{\delta^{-1}+0.5}{3\delta^{-1}+2} = 1/4 < \eta_h. \tag{65}$$

Additive representation

We need to check if there exists a sequence of h_n k_v+1-times continuously differentiable functions depending on Y_j, X_j, Z_j for observations when $M_j = 0$ such that

$$E_{Y_i,Z_i^1|M_i=1}[\hat{e}(Y_i, Z_i^1; \beta)] = n^{-1} \sum_{j:M_j=0} h_n(Y_j, X_j, Z_j) + o_p(n^{-1/2}).$$

We are following the proof of Theorem 1 in Ait-Shalia (1994) and argue that since the Hadamard-derivative

$$... \tag{66}$$

is in C^0, which means that the first order expansion

$$... \tag{67}$$

is valid by Lemma 1 of the Proof of Theorem 1 in the paper.

Smoothness h_n can take the Hadamard-derivative

$$.... \tag{68}$$

Now we need to prove that this function is k_v+1-times differentiable with bounded derivative uniformly across n. Since in the kernel function is symmetric

in X_i and X_j, the g_0 function is p-times differentiable and does not depend on n, the h_n satisfies the required smoothness assumption by the same argument we applied to show the $\hat{e}$ is $k_v + 1$-times continuously differentiable with uniformly bounded derivatives.

6.3.1 Way2

For this we consider that $\hat{e}$ is of the form

$$\hat{e}(Y_i, Z_i^1) = \sum_j \frac{Sz_{i,j}}{\hat{N}e_{i,k}}, \tag{69}$$

where $Sz_{i,j}$ is the numerator that depends on V_i, Y_j, X_j, Z_j and $\hat{N}e_{i,k}$ is the denominator that depends on all the V_k-s with $M_k = 0$ and Y_i, Z_i^1. Crucially, $|\hat{N}e_{i,k} - Ne_i| = o_p(n^{-r})$ for some power $\infty > r > 0$ and Ne_i uniformly bounded away from 0 and ∞ (Assumption 3).

Since by the definition of $\hat{e}$ it converges uniformly to e with some power rate,

$$\hat{e}(Y_i, Z_i^1) = \sum_j \frac{Sz_{i,j}}{Ne_i} \sum_{\nu=0}^{P} \left(\frac{Ne_i - \hat{N}e_{i,k}}{Ne_i} \right)^{\nu} + \int \frac{\sum_j Sz_{i,j}}{\widehat{Ne}_{i,k}} \cdot \left(\frac{Ne_i - \widehat{Ne}_{i,k}}{Ne_i} \right)^{P+1} \tag{70}$$

$$= \sum_j \frac{Sz_{i,j}}{Ne_i} \sum_{\nu=0}^{\infty} \left(\frac{Ne_i - \hat{N}e_{i,k}}{Ne_i} \right)^{\nu}, \tag{71}$$

there exists $P < \infty$ for which

$$(Ne_i - \hat{N}e_{i,k})^{P+1} = o_p\left(\sqrt{n}^{-1}\right), \tag{72}$$

which also means that $((Ne_i - \hat{N}e_{i,k})/Ne_i)^{P+1} = o_p\left(\sqrt{n}^{-1}\right)$ as well, by the boundedness properties of Ne_i and so by the Cauchy-Schwarz inequality, for a high enough finite P we get that

$$\int \frac{\sum_j Sz_{i,j}}{\widehat{Ne}_{i,k}} \cdot \left(\frac{Ne_i - \widehat{Ne}_{i,k}}{Ne_i} \right)^{P+1} f_{v|m=1}(v_i) dv_i = o_p\left(\sqrt{n}^{-1}\right), \tag{73}$$

as the first term of the integrand is bounded. This means we only need to prove

that

$$\int \frac{\sum_j Sz_{i,j}}{Ne_i} \left(\frac{Ne_i - \hat{Ne}_{i,k}}{Ne_i} \right)^{\nu} f_{v|m=1}(v_i) dv_i = o_p\left(\sqrt{n}^{-1}\right) \tag{74}$$

for $\infty > P > \nu > 0$ under our assumptions.

Now we turn to the specific estimator[4] for which the integral above is

$$\int \sum_j \frac{K[H^{-1}(v_i - V_j)] g_0(Y_j, X_j, Z_j, \beta)}{nh f_{v|m=1}(v_i)} \cdot \left(\frac{f_{v|m=0}(v_i) - \hat{f}_{v|m=0}(v_i)}{f_{v|m=0}(v_i)} \right)^{\nu} f_{v|m=0}(v_i) dv_i, \tag{75}$$

for

$$\hat{f}_{v|m=0}(v_i) = \frac{\sum_{k:M_k=0} K[H^{-1}(v_i - V_k)]}{nh}. \tag{76}$$

Let us examine only

$$\sum_j \frac{K[H^{-1}(v_i - V_j)] g_0(Y_j, X_j, Z_j, \beta)}{nh} \cdot \left(f_{v|m=0}(v_i) - \hat{f}_{v|m=0}(v_i) \right)^{\nu} = \tag{77}$$

$$= n^{-1} \sum_j \frac{K[H^{-1}(v_i - V_j)] g_0(Y_j, X_j, Z_j, \beta)}{h} \cdot \left(\frac{\sum_k h f_{v|m=0}(v_i) - K[H^{-1}(v_i - V_k)]}{nh} \right)^{\nu}$$

$$= n^{-1} \sum_j \frac{K[H^{-1}(v_i - V_j)] g_0(Y_j, X_j, Z_j, \beta)}{h} \cdot \left(\frac{h f_{v|m=0}(v_i) - K[H^{-1}(v_i - V_j)]}{nh} \right)^{\nu}$$

$$+ \sum_{\iota} \binom{\nu}{\iota} n^{-1} \sum_j \frac{K[H^{-1}(v_i - V_j)] g_0(Y_j, X_j, Z_j, \beta)}{h} \cdot \left(\frac{h f_{v|m=0}(v_i) - K[H^{-1}(v_i - V_j)]}{nh} \right)^{\nu-\iota}$$

$$\cdot \left(\frac{\sum_{k \neq j} h f_{v|m=0}(v_i) - K[H^{-1}(v_i - V_k)]}{nh} \right)^{\iota}.$$

We do not need to think about the first term, since it is a smooth function of v_i and V_j.

One may realize that after a +- empirical process term that is an additive fn of V_j only, we have either i) a degenerate U-process of order 2 with a smooth, bounded kernel, which gives the term is $o(n^{-1})$ or we have a U-process of order

[4]The argument above can be repeated for a wide range of nonparametric estimators, like local regression, series estimators etc.

at least 3, with regular properties once again, so that the term is $o(n^{-1/2})$. This proved to be a huge undertaking.

□

6.4 Proposition 2

Before we would start, we prove that the infeasible optimal weighting matrix is indeed optimal.

Now we prove Proposition 2.