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 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. 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)

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

Example 2 (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

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

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 estima-*

tor is defined with population moments

$$\left[\begin{array}{c} (1 - M_i) g_0(Y_i, X_i, Z_i; \beta)] \\ M_i e(Y_i, Z_i^1; \beta) \end{array} \right] = 0, \tag{6}$$

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

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

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)

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

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 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 X_i given Z_i, M_i as $f_{x|z,m}$. Moreover, we define notation for the Jacobians of the population mo-

ments 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 are going to 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 requiring 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* $Supp(Z_i^1) \subset \mathbb{R}^{k_1}$ *for which* $E[g_0(Y_i, X_i, Z_i, \beta)|Z_i^1, M_i = 0] = 0.$

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, and it is useful when..... 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.

The following proposition describes the asymptotic properties of the imputation GMM estimator.

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 $\tilde{\Omega} = \lim_n E[n\tilde{g}(\beta)'\tilde{g}(\beta)]$.

Under Assumption 2, given that $\sup_{y,z} E|e(y,z,b) - \hat{e}(y,z,b)|| = o_p(n^{-1/2})$ 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 $\Omega = \lim_n E[n\hat{g}(\beta, \hat{e})'\hat{g}(\beta, \hat{e})]$.

2.1.1 Convergence rate when using the Nadaraya-Watson estimator

We estimate the conditional expectation by a Nadaraya-Watson type estimator. For simplicity pf exposition, we assume that all observables are continuously distributed, and we also denote $\tilde{Z}_i = Y_i, Z_i^1$. 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.

$$\hat{e}(y_i, z_i^1; b) = \frac{\sum_j K[H^{-1}(\tilde{z}_i - \tilde{z}_j)] g_0(y_j, x_j, (z_j^0, z_j^1), b)}{\sum_j K[H^{-1}(\tilde{z}_i - \tilde{z}_j)]}. \tag{17}$$

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 $\prod h_k = h$.

Assumption 3. *Our estimation assumptions:*

a) $h_{max} \to 0$,

b) $nh_{max}^{k_1} \to \infty$,

c) K is a Parzen-Rosenblatt kernel (second order),

d) $Supp(\tilde{Z}_i)$ *is compact, with the joint pdf bounded away from zero and infinity on the support,*

e) the pdf for $\tilde{Z}_i$ *is twice differentiable,*

f) the conditional distribution function $f_{x,z^0|z^1,y}(x, y, z)$ *is twice differentiable with bounded Hessian.*

Some of these assumptions are stronger than necessary (notably, conditions 3/d) and 3/f), but they simplify the algebra greatly. We conclude that as long as the bandwidth h_{max} is $o(n^{-1/4})$, we only have to worry about the contribution of the variances, under the restriction that the $\hat{e}$ converges uniformly to the conditional expectation as a function, which gives the condition that $nh \to \infty$.

Corollary 1. *If* $\hat{e}$ *is the Nadaraya-Watson estimator with* $k_1 < 3$*, under Assumption 2-3*

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

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

The corollary suggests that we need $k_1 < 3$, if every term in the $\tilde{Z}_i$ vector is continuous (including the LHS variable). As already mentioned above, discrete dimensions of $\tilde{Z}_i$-s are allowed with simple modifications of the estimator and theory, and they do not contribute to the curse of dimensionality, so only the number of continuous conditioning variables needs to be lower than 4. There are

two reasons to include an (always observed) RHS variable into the imputation moments as conditioning variable:

- Weakening of the missing-at-completely-random assumption: we think the variable is related to missingness,
- 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 it may be related to missingness, it needs to be included in the estimator.

Example 1: Probit with Maximum Likelihood This moment is not additive, so we add the popuLation moments

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

Even though we chose the MLE moments to utilize the data points that are completely observed, we can still add the additive moments only from the previous example as imputation moments if we so wish.

2.1.2 Moments additive in X_i and Y_i

In order to preserve the simplicity of the approach of Abrevaya and Donald (2017), we may only want to focus on the moments that are additive in X_i and Y_i. If a particular row of g_0 is such for some known h_1, h_2 functions, then

$$E[M_i E[g_0(Y_i, X_i, Z_i, b)|Y_i, Z_i^1]] = E\left[M_i(h_1(Y_i, Z_i) + E[h_2(X_i, Z_i)|Z_i^1])\right], \quad (18)$$

due to[1] 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 conditional expectation may converge for higher dimension of Z_i^1 in this case ($k_1 < 4$).

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

Example 2: Probit with NLS The imputation moments we add are

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

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

$$E\left[\begin{array}{c} M_i E[X_i|Z_i^1,Y_i](Y_i - E[\Phi(aX_i+Z_ic)|Z_i^1]) \\ M_i Z_i^0(Y_i - E[\Phi(aX_i+Z_ic)|Z_i^1]) \\ M_i Z_i^1(Y_i - E[\Phi(aX_i+Z_ic)|Z_i^1]) \end{array}\right], \tag{20}$$

which are more simple to estimate. Abrevaya and Donald (2017) only uses the additive moments in the case when the conditional expectation is further parametrized to ensure it can be estimated $\sqrt{n}$-consistently. To preserve the appealing simplicity of their approach, we will also restrict our attention to these moments in the continuation of the example.

2.2 The role of the weighting matrix and efficiency

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

$$\begin{aligned}(\beta_n-\beta)(\beta_n-\beta)' &= ((\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{g}_1)((\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{g}_1)' = \\ &= (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{\Omega}_0\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}.\end{aligned} \tag{21}$$

Now let us set

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

then we get that

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

We note that

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

which means this is the infeasible optimal weighting for large samples. This

optimal weighting matrix can be estimated by the inverse of $n^{-1}\sum gg'$, which is a block-diagonal matrix.

$$\hat{\Omega} = \begin{bmatrix} \hat{\tilde{\Omega}} & 0 \\ 0 & \hat{B} \end{bmatrix}. \tag{23}$$

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 ($\hat{B}$) is positive definite if the additional moments do not have a zero optimal weight as n tends to infinity. In this case $diag(G\hat{W}G)^{-1}$ is smaller or equal than the diagonal of the optimal covariance matrix of the estimator that does not contain the added moments (which is $(\tilde{G}'^{-1}\hat{\tilde{\Omega}}\tilde{G})$).

Proposition 2. *Under Assumption 2 and if $\hat{B}/||\hat{\Omega}||$ is bounded eventually,*[2] *$MSE(\beta_n) \leq MSE(\tilde{\beta}_n)$ ev. for 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 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 1. 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, this inference introduces additional noise and loss of degrees of freedom in finite samples, so if the applied researcher implements the imputation method for too many dimensions, imputation will slightly *increase* standard errors in finite samples.

Remark 2. 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').

[2] 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." sometimes.

3 Monte Carlo simulation

4 Conclusion

5 References

Abrevaya and Donald (2017) Ichimura and Newey (2015) Pakes and Pollard (1989) Chernozhukov et al. (2018)

6 Appendix

6.1 Proposition 1

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

$$\begin{aligned} 0 &= \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{g}(\hat{\beta}_n;\hat{e}) = \\ &= \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{g}(\beta;\hat{e}) + \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{G}(\bar{\beta}_n;\hat{e})(\hat{\beta}_n - \beta) = \\ &= \hat{G}'\hat{W}\hat{g}_1 + \hat{G}'\hat{W}\bar{G}(\hat{\beta}_n - \beta). \end{aligned} \tag{24}$$

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 $\bar{\beta}$. We also introduce

$$\hat{g}_1 = \hat{g}(\beta;\hat{e}), \tag{25}$$

$$g_1 = g(Y_i, X_i, Z_i, M_i;\beta;e). \tag{26}$$

Given Assumption **??**, we can prove that $\hat{\beta}_n$ consistently estimates β, as $||\hat{G}'\hat{W}\bar{G}||$ is bounded and $\hat{g}_1 \to E[g_1] = 0$ with probability approaching 1. This in turn yields

$$\begin{aligned} \hat{\beta}_n - \beta &= -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\hat{g}_1 \\ &= -(G'WG)^{-1}G'W\hat{g}_1 + o_p(\hat{g}_1). \end{aligned} \tag{27}$$

6.2 Corollary 1

$$\hat{E}(zy|z=z_i) = (nh)^{-1}\frac{\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i)}{(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]}, \tag{28}$$

where the denominator clearly converges in probability to $f(z_i)$, uniformly, so we are going to ignore it, and focus on the expected value of

$$(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[zg(\alpha x+\beta z)|z=z_i]. \tag{29}$$

First, let us calculate

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i)|\mathbf{z}] = \\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i E[g(\alpha x_j+\beta z_i)|\mathbf{z}] = \\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i \int g(\alpha x+\beta z_i) f_{x|z}(x,z_j)dx, \end{aligned} \tag{30}$$

which gives

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[z_i g(\alpha x+\beta z_i)|z_i]|\mathbf{z}] = \\ &= (nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i \int g(\alpha x+\beta z_i)(f_{x|z}(x,z_j)-f_{x|z}(x,z_i))dx. \end{aligned} \tag{31}$$

It is interesting that it is only the conditional distribution that has the discrepancy. Taking now expectation w.r.t. z_j as well,

$$\begin{aligned} &E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[z_i g(\alpha x+\beta z_i)|z_i]|z_i] = \\ &= h^{-1}\int K[H^{-1}(z_i-z)]z_i \int g(\alpha x+\beta z_i)(f_{x|z}(x,z)-f_{x|z}(x,z_i))f(z)dxdz = \\ &= \int K[\Delta z](z_i \int g(\alpha x+\beta z_i) Df_{x|z}(x,z_i)dx\Delta z(f(z_i)+Df(\bar{z})\cdot\Delta z\cdot H)d\Delta z + \\ &+ \int \Delta z' H D^2 f_{x|z}(x,\bar{\bar{z}})H\Delta z dx \end{aligned} \tag{32}$$

after taking a second-order Taylor expansion in $f(z)$ and estimating $f_{x|z}(x,z) - f_{x|z}(x,z_i)$ similarly, finally, substituting $\Delta z = H^{-1}(z_i - z)$ for integration. By

our boundedness assumptions, this is going to be bounded uniformly over z_i.

Given that we have second order kernel, we collect the terms and take integrals (everything has the same rate uniformly over z_i)

$$E[(nh)^{-1}\sum_j K[H^{-1}(z_i - z_j)]z_i g(\alpha x_j + \beta z_i) - E[z_i g(\alpha x + \beta z_i)|z_i]] = \quad (33)$$
$$= O(h_{max}^2)$$

(rewrite this), but checked

6.3 Proposition 2

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

Now we prove Proposition 2.

6.4 Additive g_0

$$\begin{aligned}
&E[M_i E[g_0(Y_i, X_i, Z_i, b)|Y_i, Z_i^1]] = \quad (34)\\
&= E[M_i E[h_1(Y_i, Z_i) + h_2(X_i, Z_i)|Y_i, Z_i^1]] =\\
&= E[M_i E[E[h_1(Y_i, Z_i) + h_2(X_i, Z_i)|Y_i, Z_i]|Y_i, Z_i^1]] =\\
&= E[M_i h_1(Y_i, Z_i)] + E[M_i E[h_2(X_i, Z_i)|Y_i, Z_i^1]] =\\
&= E[M_i h_1(Y_i, Z_i)] + E[E[M_i E[h_2(X_i, Z_i)|Y_i, Z_i^1]|Z_i^1]] =\\
&= E[M_i h_1(Y_i, Z_i)] + E[E[M_i|Z_i^1]E[h_2(X_i, Z_i)|Z_i^1]] =\\
&= E[M_i h_1(Y_i, Z_i)] + E[P[M_i = 1|Z_i^1]E[h_2(X_i, Z_i)|Z_i^1]] =\\
&= E[M_i h_1(Y_i, Z_i)] + E[M_i E[h_2(X_i, Z_i)|Z_i^1]] =\\
&= E\left[M_i(h_1(Y_i, Z_i) + E[h_2(X_i, Z_i)|Z_i^1])\right]
\end{aligned}$$