Simple nonlinear imputation

December 2020

1 Introduction

In this paper we consider GMM estimation in models where given a known function g_0, the LHS variable (Y_i) and RHS variables ($X_i \in \mathbb{R}, Z_i \in 1 \times \mathbb{R}^k$) there is a set of differentiable population moment functions g_0 that are zero in expectation if and only if they are evaluated at the true parameter values (β),

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

where the vector $\beta \in \mathbb{R}^p$ is what the researcher aims to estimate.

The key problem in our setting is that the scalar X_i variable is missing whenever M_i, the missingness indicator is 1. We are exploring the properties of a simple imputation GMM estimator under the further assumption that the conditional distribution of X_i and Y_i given Z_i is independent of M_i.

Example 1 (Probit Non-linear Least Squares). 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. The following model connects the observables:

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

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

In the following we denote the cumulative distribution function of the standard normal distribution by Φ and the probability density function by ϕ. One way of estimating the model is non-linear least squares, when the set of population moments is

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

Example 2 (Probit Maximum Likelihood)*.* Assume that the data generating process is the same as in the previous example (equations (2)-(3)). Another way to estimate the coefficients is based on the likelihood principle that corresponds to the population moment

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

In both examples, our key exclusion restrictions are satisfied when the missingness is caused by the always observed variables Z_i, but conditional on these, Y_i and X_i are not related to it.

Given our setup, it is possible to estimate the parameter of interest using the completely observed records in the data based on the population moment g_0 (the complete case GMM estimator). We show when it is advantageous to augment the original moments using simple imputation to increase the efficiency of the estimates.

LITERATURE REVIEW, SELLING MISSING HERE

2 A simple imputation GMM estimator

2.1 Model assumptions and estimators

We denote the conditional pdf of X_i given Z_i, M_i as $f_{x|z,m}$. Let us denote the set of random vectors with the same support as $Supp(X_i, Z_i)$ as $\mathcal{W}$.

Assumption 1 (Model)**.** *There exists a known differentiable g_0 vector-valued function such that*

1. $E[g_0(Y_i, X_i, Z_i, b)] = 0 \iff b = \beta$*,*
2. $E[g_0(Y_i, X_i, Z_i, b)|X_i, Z_i, M_i] = 0 \Leftarrow b = \beta$*,*
3. $P[M_i = 1|X_i, Z_i] < 1 \; X_i, Z_i - a.s.$

4. *Z_i can be partitioned into $Z_i = [Z_i^0, Z_i^1]$, such that conditional on Z_i^1, the distribution of X_i, Z_i^0 is the same whether $M_i = 0$ or $M_i = 1$.*

The first condition in Assumption 1 ensures identification in the case without missingness. The second and third conditions are the usual assumptions for the validity of the *complete case GMM* estimator[1] under missing-at-random (MAR), when the researcher simply omits the observations with missing values. The fourth condition is the weakened version of the missing-at-completely-random assumption often assumed by researchers, but it is a stronger assumption than MAR. These conditions are satisfied if the (Y, X_i, Z_i^0) vector is independent of M_i, conditional on Z_i^1. Further we denote the dimension of the support of Z_i^1 by k_1.

Next we define three GMM estimators:

1. The infeasible *full-data GMM estimator* with the moment

$$E[g_0(Y_i, Z_i, X_i; \beta)] = 0,$$

2. The *complete case GMM estimator* that is based on the population moment

$$E[\tilde{g}(Y_i, Z_i, X_i, M_i; \beta)] = E\left[(1 - M_i)g_0(Y_i, X_i, Z_i; \beta)\right] = 0 \ a.s. \tag{4}$$

   for $\tilde{g} : Supp(Y_i, Z_i, X_i, M_i) \times B$ for $B \subset \mathbb{R}^p$,

3. The *imputation GMM estimator* with population moment

$$E[g(Y_i, Z_i, X_i, M_i; b; e)] = \left[\begin{array}{c} (1 - M_i)g_0(Y_i, X_i, Z_i, b)] \\ M_i e(Y_i, Z_i^1; b) \end{array} \right], \tag{5}$$

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

$$e(y, z^1, b) = E[g_0(Y_i, X_i, Z_i, b)|Y_i = y, Z_i^1 = z^1]. \tag{6}$$

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

$$\hat{g}_0(y_i, z_i, x_i; b) = n^{-1} \sum_{i=1}^{n} g_0(y_i, x_i, z_i, b) \tag{7}$$

[1] See the definition below.

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

$$\hat{Q}_n^0(b) = \hat{g}_0(b)'\hat{W}_0\hat{g}_0(b) \tag{8}$$

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 based on the moment

$$\begin{aligned}\hat{\tilde{g}}(b) &= n^{-1}\sum_{i=1}^{n}\tilde{g}(y_i, z_i, x_i, m_i; b) = \\ &= n^{-1}\sum_{i=1}^{n}(1-m_i)g_0(y_i, z_i, x_i; b),\end{aligned} \tag{9}$$

and defined as the M-estimator minimizing

$$\hat{\tilde{Q}}_n(b) = \hat{\tilde{g}}(b)'\hat{\tilde{W}}\hat{\tilde{g}}(b) \tag{10}$$

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

$$\hat{\tilde{W}} \xrightarrow{p} \tilde{W}. \tag{11}$$

Under Assumption 1 this estimator is consistent.

In addition to the feasible identifying moments $\tilde{g}$ we also add imputation moments to our *imputation GMM estimator* in order to increase efficiency. The function g has the same first arguments as $\tilde{g}$, while the last argument is a function that represents the conditional expectation of $g_0(Y_i, X_i, Z_i, b)$ given $Z_i = z$ and $Y_i = y$. Clearly, this function is identified from the complete subsample. Define the sample analogues (for a sample of size n)

$$\begin{aligned}\hat{g}(b, \hat{e}) &= n^{-1}\sum_{i=1}^{n}g(y_i, z_i, x_i, m_i; b; \hat{e}) = \\ &= n^{-1}\sum_{i=1}^{n}\left[\begin{array}{c}(1-m_i)g_0(y_i, z_i, x_i; b) \\ m_i\hat{e}(y_i, z_i^1; b))\end{array}\right],\end{aligned} \tag{12}$$

where $\hat{e}$ is an estimator of the conditional expectation e. At the end of this section we give the Nadaraya-Watson estimator as a specific example for a viable

$\hat{e}$. The *imputation GMM estimator* is minimizing

$$\hat{Q}_n(b) = \hat{g}(b;\hat{e})'\hat{W}\hat{g}(b;\hat{e}) \tag{13}$$

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

$$\hat{W} \xrightarrow{p} W. \tag{14}$$

2.2 Asymptotic properties

In the following arguments we closely follow Ichimura and Newey (2015) and Chernozhukov et al. (2018). We denote the imputation estimator as a random variable by $\hat{\beta}_n$, while the true value by β. We also introduce the notation $\hat{G}(\beta;\hat{e})$ for the derivative of $\hat{g}$ with respect to the parameter vector. Correspondingly, the derivative of $E[g]$ evaluated at β is denoted by G.

Assumption 2. *We make several regulatory assumptions.*

- g_0 *is continuously differentiable with bounded derivatives,*
- *random sampling,*
- $\beta \in B$, *a compact set,*
- $G'WG$ *is invertible a.s.,*
- $\hat{e}(y, z^1, b) \xrightarrow{p} e(y, z^1, b)$ *uniformly,*
- $\sup_{y,z^1} \hat{e}(y, z^1, b) < \infty$ *almost surely for all* $b \in B$,

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

$$\begin{aligned} 0 &= \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{g}(\hat{\beta}_n;\hat{e}) = \\ &= \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{g}(\beta;\hat{e}) + \hat{G}'(\hat{\beta}_n;\hat{e})\hat{W}\hat{G}(\bar{\beta}_n;\hat{e})(\hat{\beta}_n - \beta) = \\ &= \hat{G}'\hat{W}\hat{g}_1 + \hat{G}'\hat{W}\bar{G}(\hat{\beta}_n - \beta). \end{aligned} \tag{15}$$

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

$$\hat{g}_1 = \hat{g}(\beta; \hat{e}), \tag{16}$$

$$g_1 = g(Y_i, X_i, Z_i, M_i; \beta; e). \tag{17}$$

Given Assumption 2, 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

$$\begin{aligned}\hat{\beta}_n - \beta &= -(\hat{G}'\hat{W}\bar{G})^{-1}\hat{G}'\hat{W}\hat{g}_1 \\ &= -(G'WG)^{-1}G'W\hat{g}_1 + o_p(\hat{g}_1).\end{aligned} \tag{18}$$

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

Proposition 1. *Under Assumptions 1-2, if* $\sup_{y,z} E|e(y,z,b) - \hat{e}(y,z,b)|| = o_p(n^{-1/2})$*, then*

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

with $\Omega = \lim E[\hat{g}_1'\hat{g}_1]$.

We also note that the complete case GMM estimator has the analogous properties under our assumptions.

Convergence rate of $\hat{e}(y, z^1, b)$ *for the Nadaraya-Watson estimator.* We estimate the conditional expectation by a Nadaraya-Watson type estimator. For the sake of simplicity, we assume that all observables are continuously distributed, and we also denote $\tilde{Z}_i = Y_i, Z_i^1$. We explore the case when the LHS variable is discrete, which makes the estimation of the conditional expectation nuisance parameter less complicated.

$$\hat{e}(y_i, z_i^1; b) = \frac{\sum_j K[H^{-1}(\tilde{z}_i - \tilde{z}_j)]g_0(y_i, x_j, (z_j^0, z_i^1), b)}{\sum_j K[H^{-1}(\tilde{z}_i - \tilde{z}_j)]}. \tag{19}$$

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:*

1. $h_{max} \to 0$
2. $nh_{max}^{k_z} \to \infty$
3. *K is a Parzen-Rosenblatt kernel (second order)*
4. *$Supp(\tilde{Z}_i)$ is compact, with the strong pdf assumption*
5. *the pdf for Z_i is twice differentiable*
6. *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 number 4 and 6). 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 $nh \to \infty$. This implies that we need that the $k_1 < 3$ that is involved in the calculations of the conditional expectations. In addition, we note that discrete $\tilde{Z}_i$-s are allowed with simple modifications of the estimator and theory.

Corollary 1. *If $\hat{e}$ is the Nadaraya-Watson estimator with $k_1 < 3$, under Assumption 1-3 the imputation GMM estimator $\hat{\beta}_n$ is such that*

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

with $\Omega = \lim E[\hat{g}_1' \hat{g}_1]$.

The corollary suggests that the dimensionality of Z_i included in the conditioning is crucial for imputation to work. There are two reasons to include an (always observed) RHS variable into the imputation moments:

- Weakening of the missing-at-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.

Next we are going to analyze our two examples and also show how with certain type of added imputation moment we can decrease the effective number of dimensions we need to condition on. This also connects this paper with the original strategy of Abrevaya and Donald (2017).

2.3 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

$$\begin{aligned}
&E[M_iE[g_0(Y_i, X_i, Z_i, b)|Y_i, Z_i^1]] = \\
&\quad = E[M_iE[h_1(Y_i, Z_i) + h_2(X_i, Z_i)|Y_i, Z_i^1]] = \\
&\quad = E[M_iE[E[h_1(Y_i, Z_i) + h_2(X_i, Z_i)|Y_i, Z_i]|Y_i, Z_i^1]] = \\
&\quad = E[M_ih_1(Y_i, Z_i)] + E[M_iE[h_2(X_i, Z_i)|Y_i, Z_i^1]] = \\
&\quad = E[M_ih_1(Y_i, Z_i)] + E[E[M_iE[h_2(X_i, Z_i)|Y_i, Z_i^1]|Z_i^1]] = \\
&\quad = E[M_ih_1(Y_i, Z_i)] + E[E[M_i|Z_i^1]E[h_2(X_i, Z_i)|Z_i^1]] = \\
&\quad = E[M_ih_1(Y_i, Z_i)] + E[P[M_i = 1|Z_i^1]E[h_2(X_i, Z_i)|Z_i^1]] = \\
&\quad = E[M_ih_1(Y_i, Z_i)] + E[M_iE[h_2(X_i, Z_i)|Z_i^1]] = \\
&\quad = E\left[M_i(h_1(Y_i, Z_i) + E[h_2(X_i, Z_i)|Z_i^1])\right]
\end{aligned} \tag{20}$$

due to 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 ($k_1 < 4$).

Example 1: Probit with NLS The imputation moments we add are

$$E\begin{bmatrix} M_iE[X_i|Z_i^1, Y_i](Y_i - E[\Phi(aX_i + Z_ic)|Z_i^1, Y_i]) \\ M_iE[Z_i^0|Z_i^1, Y_i](Y_i - E[\Phi(aX_i + Z_ic)|Z_i^1, Y_i]) \\ M_iZ_i^1(Y_i - E[\Phi(aX_i + Z_ic)|Z_i^1, Y_i]) \end{bmatrix}. \tag{21}$$

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

$$E\left[\begin{array}{c} M_iE[X_i|Z_i^1,Y_i](Y_i-E[\Phi(aX_i+Z_ic)|Z_i^1]) \\ M_iZ_i^0(Y_i-E[\Phi(aX_i+Z_ic)|Z_i^1]) \\ M_iZ_i^1(Y_i-E[\Phi(aX_i+Z_ic)|Z_i^1]) \end{array}\right], \tag{22}$$

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.

Example 2: Probit with Maximum Likelihood This moment is not additive, so we add the popuLation moments

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

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.

2.4 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

$$\begin{aligned}&(\hat\beta_n-\beta)(\hat\beta_n-\beta)' \\ &\quad=((\hat G'\hat W\hat G)^{-1}\hat G'\hat W\hat g_1)((\hat G'\hat W\hat G)^{-1}\hat G'\hat W\hat g_1)'= \\ &\quad=(\hat G'\hat W\hat G)^{-1}\hat G'\hat W\hat\Omega_0\hat W\hat G(\hat G'\hat W\hat G)^{-1}.\end{aligned} \tag{23}$$

Now let us set

$$\hat W^{-1}=\hat\Omega_0,$$

then we get that

$$(\hat\beta_n-\beta)(\hat\beta_n-\beta)'=\left(\hat G'\hat\Omega_0\hat G\right)^{-1}. \tag{24}$$

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.

$$\hat{\Omega}=\begin{bmatrix}\hat{\tilde{\Omega}} & 0\\ 0 & \hat{B}\end{bmatrix}. \tag{25}$$

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 1-2 and if $\hat{B}/||\hat{\Omega}||$ is bounded eventually, $MSE(\hat{\beta}_n)\leq MSE(\tilde{\beta}_n)$ ev. for any admissible weighting of the $\tilde{\beta}_n$ estimator. The inequality is 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 these 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 estimates nuisance parameter, its relative weight is set to be arbitrarily close to zero, eventually. The optimal weighting matrix selects the moments automatically so that the estimator is always $\sqrt{n}$-consistent with an asymptotic variance-covariance matrix that is the same as for the complete case GMM estimator.
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 $k^1\geq 4$, 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’).

3 Monte Carlo simulation

4 Application

5 Conclusion

6 References

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

7 Appendix

7.1 Proposition 1

Theorem 7 checking the assumptions?

7.2 Corollary 1

$$\hat{E}(zy|z=z_i) = (nh)^{-1}\frac{\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j + \beta z_i)}{(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]}, \tag{26}$$

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

$$(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j + \beta z_i) - E[zg(\alpha x + \beta z)|z=z_i]. \tag{27}$$

First, let us calculate

$$\begin{aligned}
&E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i)|\mathbf{z}] = \\
&=(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i E[g(\alpha x_j+\beta z_i)|\mathbf{z}] = \\
&=(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i \int g(\alpha x+\beta z_i) f_{x|z}(x,z_j)dx,
\end{aligned} \tag{28}$$

which gives

$$\begin{aligned}
&E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[z_i g(\alpha x+\beta z_i)|z_i]|\mathbf{z}] = \\
&=(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i \int g(\alpha x+\beta z_i)(f_{x|z}(x,z_j)-f_{x|z}(x,z_i))dx.
\end{aligned} \tag{29}$$

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

$$\begin{aligned}
&E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[z_i g(\alpha x+\beta z_i)|z_i]|z_i] = \\
&=h^{-1}\int K[H^{-1}(z_i-z)]z_i \int g(\alpha x+\beta z_i)(f_{x|z}(x,z)-f_{x|z}(x,z_i))f(z)dxdz = \\
&=\int K[\Delta z](z_i \int g(\alpha x+\beta z_i) D f_{x|z}(x,z_i)dx\Delta z(f(z_i)+Df(\bar{z})\cdot\Delta z\cdot H)d\Delta z + \\
&+\int \Delta z' H D^2 f_{x|z}(x,\bar{\bar{z}})H\Delta z dx
\end{aligned} \tag{30}$$

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)

$$\begin{aligned}
&E[(nh)^{-1}\sum_j K[H^{-1}(z_i-z_j)]z_i g(\alpha x_j+\beta z_i) - E[z_i g(\alpha x+\beta z_i)|z_i]] = \\
&=O(h_{max}^2)
\end{aligned} \tag{31}$$

(rewrite this), but checked

7.3 Proposition 2

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

Now we prove Proposition 2.