Simple nonlinear imputation

December 2020

1 Monte Carlo simulation

We are studying a data generating process and estimator from Example 2,

$$Z_i \sim U[-2,2]^{k_z} \tag{1}$$
$$X_i = Z_i\delta_x + \epsilon_i^x \tag{2}$$
$$M_i = \mathbf{1}[|Z_i\delta_m + \epsilon_i^m| > 0.8] \tag{3}$$
$$\epsilon_i^{m,x} \sim N[0,1] \text{ i.i.d, conditional on } X_i, Z_i \tag{4}$$
$$Y_i = \mathbf{1}[\alpha X_i + Z_i\gamma \geq U_i] \tag{5}$$
$$U_i \sim N[0,1] \text{ i.i.d, conditional on } X_i, Z_i, \tag{6}$$

in addition, the normal disturbance terms $U_i, \epsilon_i^m, \epsilon_i^x$ are independent of each other as well. The coefficient vector of interest (α, γ) is estimated by NLLS, which means that

$$g_0 = E\left[\begin{array}{c} X_i[Y_i - \Phi(\alpha X_i + Z_i\gamma)] \\ Z_i[Y_i - \Phi(\alpha X_i + Z_i\gamma)] \end{array}\right]$$

as we mentioned in section **??**. We also estimate the conditional probability density function of X_i at 400 grid points on $]-2,2[$ for every Z_i vector we observe when $M_i = 1$ by

$$\hat{f}_{x|z}(x,z) = \frac{\sum_j \prod_{k=1}^{k_z} K[(z_j^k - z^k)/h_{2,n}] \cdot K[(x_j - x)/h_{1,n}]}{\sum \prod_{k=1}^{k_z} K[(z_j^k - z^k)/h_{2,n}]}, \tag{7}$$

n	Infeasible GMM	Complete case GMM	Imputation GMM
1,000			
Mean	1.013, 0.504, -2.023	1.028, 0.509, -2.046	1.037, 0.526, -2.084
St. dev.	0.111, 0.095, 0.145	0.163, 0.142, 0.212	0.466, 0.265, 0.697
MSE	0.0144	0.0316	0.0268
2,000			
Mean	1.005, 0.505, -2.011	1.011, 0.508, -2.023	1.012, 0.514, -2.036
St. dev.	0.077, 0.065, 0.099	0.110, 0.096, 0.145	0.111, 0.075, 0.121
MSE	0.0067	0.0144	0.0114
4,000			
Mean	1.003, 0.502, -2.005	1.006, 0.503, -2.010	1.006, 0.507 -2.019
St. dev.	0.054, 0.046, 0.069	0.076, 0.069, 0.099	0.076, 0.052, 0.083
MSE	0.0033	0.0068	0.0053
8,000			
Mean	1.001, 0.501, -2.003	1.004, 0.502, -2.007	1.004, 0.503, -2.010
St. dev.	0.038, 0.032, 0.049	0.055, 0.048, 0.070	0.055, 0.036, 0.058
MSE	0.0016	0.0034	0.0026

Table 1: Monte Carlo simulations for $k_v = 2$ for the NLLS estimator in the probit model with 5,000 repetitions. The true values are $[1, 0.5, -2]]$, where the first coefficient corresponds to the missing variable X_i. Besides the means and standard deviations of the estimates we included the Mean Squared Error (MSE) once averaged over the three coefficient estimates.

where K is the Gaussian kernel, and h_1, h_2 bandwidth decrease with a -0.33 rate. After this, we define

$$\hat{e}(z_i; a, c) = z_i \left(y_i - \sum_{g=1}^{400} \Phi\left(a x_g + z_i c\right) \hat{f}_{x|z}(x_g, z_i) \right), \tag{8}$$

where x_g represent the gth point on the 400-grid we chose for the support of X_i.

This estimator has the same properties as the Nadaraya-Watson estimator for $\hat{e}$, but it has the added benefit that we do not have to recalculate the non-parametric part for every function evaluation in the numerical optimization procedure. It also turns out to have nicer finite sample performance. This way the code is close to $O(n)$ computational complexity in the range of sample sizes we considered in this simulation (rather than $O(n^2)$). For the sake of simplicity, we only target the γ coefficients and use the additive results in section **??**, while omitting the first moment from g_0, corresponding to X_i.

Table 1 gives the result for $k_z = 2$ with $\alpha = 1, \gamma = (0.5, -2)$ and 5,000 iterations. We can see how the imputation estimator has an overall better performance for the γ coefficients compared to the complete case estimator. The bias is slightly higher for the imputation estimator, due to the first stage, especially for the lower sample sizes, while the variance is much lower. The two estimator gives virtually the same results for the α coefficient. The estimators are $\sqrt{n}$-consistent as prescribed above $n = 2,000$. As for numerical stability, the BFGS algorithm with imputation estimator had trouble to find the minimum for $n = 1,000$ in one occasion, and resulted in an implausibly high coefficient estimate. We included this estimate value in the mean and standard deviation calculations, but excluded the iteration when evaluating the MSE.

NOTES:

- I found that directly estimating the conditional expectation of the moments is less computationally advantageous than doing the numerical integration above (which is somewhat unexpected). In addition, the second, seemingly more complicated option also has a better finite sample performance, it seems. The cost: need to choose 2 bandwidths instead of 1. Should I rewrite the estimation section to reflect the estimator here, or is it clear that this is really just a very similar approach?
- Is it ok to leave out the first moment from g_0 here? It is more compatible with Abrevaya and Donald + simpler. (Although technically speaking, I think you can make other simple arguments to estimate $E[Y_i X_i]$ separately (just conditioning on Y_i).
- Numerical instability: I haven't programmed the evaluation of the Jacobian and the Hessian into the code. My conjecture is that if this issue is due to numerical problems, this would help. The real question is: Is this a problem? Should we improve the code in this direction? Even if the answer is 'no', we may want to talk about how to communicate this issue.
- The simulation above ran for 36 hours with 10 i5 cores and 16 GB memory. This can get somewhat lower if the Jacobian + Hessian are evaluated. I expect that a single node on the cluster at least thirds computation time.
- As it is written, the code is efficient for these lower sample sizes. The first stage is using a nice little function to do matrix manipulation, but this function is going to switch from $O(n)$ to $O(n^2)$ complexity (and needs

much more memory, for some reason) above appx. 10,000 observations. This means that scaling to n=16,000 would have be more costly.

- Currently running "what happens" when $k_z = 5$ and the bandwidth is decreasing with a -0.33 rate, regardless. More problem with numerical stability at 1,000 already seen,[1] once the results are "cleaned" from suspiciously high/low values, the MSE for the imputation estimator is still slightly below the complete case estimator (but really only slightly). I would have expected slightly above, but this may be simulation error? (Also, 1,000 seemed to be still before the point asymptotics would start to kick in.) Including 3 more Z_i-s increased computation time, and the run will take around 48 hours, even after excluding the $n = 8,000$ sample size (which can only run with 4 cores and 16 GBs on my computer and the 5,000 iterations would take around 100 hrs - this probably can be sped up 7fold for the cluster though). All in all, a wishlist should be compiled for simulations. What do you wish for? :)

[1] The complete case estimator has some problems like this. The number of excluded iterations is around 200 this time, though.