GMM UNDER MODERATELY HIGH DIMENSIONS

ABSTRACT. Here we discuss the issues that arise in GMM under increasing dimension. This is not an easy material.

1. INTRODUCTION

For the moment function

$$g(\theta) := \mathrm{E}g(X, \theta),$$

we assume that the true parameter value of interest $\theta_0 \in \Theta \subset \mathbb{R}^{d_\theta}$ satisfies:

$$g(\theta_0) = 0.$$

We have data $\{X_i\}_{i=1}^n$, which are identical copies of X, and, as a leading case we assume that they are also independent (i.i.d.). We form the empirical moment function:

$$\hat{g}(\theta) = \mathbb{E}_n g(X_i, \theta).$$

Then our estimator $\hat{\theta}$ of θ_0 is

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \ \hat{g}(\theta)' \hat{A} \hat{g}(\theta), \tag{GMM}$$

where $\hat{A}$ is a positive-definite matrix, possibly data-dependent, that converges to a non-stochastic positive-definite matrix A. We claimed that under some conditions

$$\hat{\theta} \overset{\text{a}}{\sim} N(\theta_0, V/n).$$

> To what extent does this result hold when the dimension of θ is high or when the number of moments considered is high?

2. GMM PROBLEMS THAT SEPARATE IN MANY LOW-DIMENSIONAL GMM PROBLEMS

One case that is particularly nice is when the GMM problem can be disintegrated into many low dimensional problems, and we treat each in the limited information framework. For example, when we ran distribution regressions at multiple threshold values, the dimension of all the parameters stacked together was quite high. Was this a problem? No, it turns out that we can handle very many low-dimensional problems in this way. Formally, the number of problems p could be very high but with the mild requirement that:

$$(\log p)^7 / n \to 0. \tag{2.1}$$

Formally, suppose that $\theta = (\theta_1', ..., \theta_p')'$, where each θ_j is a low dimensional parameter, namely $\dim \theta_j$ is bounded as $n \to \infty$, for a GMM problem indexed by $j = 1, ..., p$. Assume that the GMM estimator $\hat{\theta} = (\hat{\theta}_1', ..., \hat{\theta}_p')'$ obeys the asymptotic linearization property:

$$\sqrt{n}(\hat{\theta}_j - \theta_j) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\varphi_j(X_i)}_{\text{influence function}} + \underbrace{r_j}_{\text{remainder}},$$

where the remainder can be shown to be small

$$\max_{j \leq p} \|r_j\| = o_p(1/\log p).$$

Conditions for such linearization are weak. Recall that for GMM, the influence function is

$$\varphi_j(X_i) = -(G_j' A_j G_j)^{-1} G_j' A_j g_j(X_i, \theta_{0j}),$$

where the index j signifies the dependence of the usual quantities on the j-th problem. Conditions on the remainder terms can also be established under mild conditions.

Then it turns out that under some technical moment conditions and (2.1) we can still claim that

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{\sim} N(0, V), \tag{2.2}$$

but in the sense that

$$\sup_{A \in \mathcal{R}} \left| \mathrm{P}(\sqrt{n}(\hat{\theta} - \theta_0) \in A) - \mathrm{P}(N(0, V) \in A) \right| \to 0,$$

where $\mathcal{R}$ is the collection of rectangular sets in $\mathbb{R}^{\sum_{j=1}^{p} \dim \theta_j}$. The distinction that it holds for rectangular sets is important in very high dimensions, as the result is not true for elliptical regions or other convex regions.

For inference purposes, we can also employ bootstrap or at least quick bootstrap, where we bootstrap the estimated influence functions. These results follow from central limit theorems and bootstrap results for high dimensional vector means or approximate means developed in [3].

3. GMM problems that do not separate into low-dimensional problems

We study now the situation where the number of parameters or the number of moments is large. This case includes nonlinear least squares or M-estimation problems with many regressors, and the Arellano-Bond estimator for linear panel models with sequential exogeneity and large T.

We partition the parameter vector as:

$$\theta = (\alpha', \gamma')', \quad \theta_0 = (\alpha_0', \gamma_0')',$$

where α is the target parameter and γ is the nuisance parameter, and α_0 and γ_0 are their respective true values.

The dimension of α is low, the dimension of γ is high. We approximate this situation as $p = \dim(\gamma) \to \infty$ as $n \to \infty$, while $d = \dim(\alpha)$ is fixed. Also the number of moments used in GMM, $m = \dim(g(X_i, \theta_0))$, could be high, so we can approximate this situation as $m \to \infty$ as $n \to \infty$.

Assertion 1 (GMM Inference with Moderately Many Parameters and Moments)**.** *There exist regularity conditions such that if the square of the dimensionality of the nuisance parameter and the number of moments is small compared to the sample size, namely that:*

$$(p+m)^2/n \to 0 \text{ as } n \to \infty,$$

then the approximate normality and consistency results continue to hold:

$$\sqrt{n}(\hat{\alpha} - \alpha) \overset{a}{\sim} N(0, V_{11}),$$

where V_{11} is the upper-left block of V in (2.2)*.*

Remark 1. Sufficient conditions are given, for example, by [7] for GMM problems with $m \to \infty$ and p is fixed; and by [5, 4] for nonlinear panel data models where $m \propto p \to \infty$. For exactly identified exogeneous linear models, this condition can often be improved to requiring that the dimension is small compared to the sample size, $p/n \to 0$. For exogenous linear models, very strong results were obtained in a sequence of papers by [6, 1, 2], which also cover the case where $p/n \to c > 0$, which gives rise to additional terms in the variance formula. ■

The result above has a simple practical message: in nonlinear models or models with endogeneity, p^2 and m^2 should be small compared to n. For example, we saw that in Arellano-Bond approach the number of moments $m = O(T^2)$, which could be too large relative to the sample size nT. This seems to be binding in the second empirical example that we considered where $n = 147$ and $T = 19$. This practical rule may be too rough as a guide in some applications. In such cases we may carry out Monte-Carlo experiments, using data-generating processes that mimic the empirical settings one is facing, to see how the methods perform.

To understand Assertion 1, let us focus on the case where $p = m$. An asymptotic second order expansion of $\hat{\alpha}$ around α_0 gives

$$\hat{\alpha} - \alpha_0 = Z_n/\sqrt{n} + b/n + r_n,$$

where $Z_n \overset{a}{\sim} N(0, V_{11})$, $b = O(p)$ is a first order bias term, and r_n is the higher order remainder such as $r_n = O_p((p/n)^{3/2} + p^{1/2}/n)$. Then,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \overset{a}{\sim} N(0, V_{11})$$

if both

$$\sqrt{n}b/n \to 0, \quad \text{i.e. } p^2/n \to 0,$$

and

$$\sqrt{n}r_n \to_P 0, \quad \text{i.e. } p^{3/2}/n \to 0.$$

Example 1 (A Contrived Example)**.** Suppose that we are interested in the parameter $\alpha = \gamma'\gamma$, and $\alpha_0 = \gamma_0'\gamma_0 = 1$. Suppose also that $\hat{\gamma} - \gamma_0 = G_n \sim N(0, I_p/n)$. Then the $\hat{\alpha} = \hat{\gamma}'\hat{\gamma}$ obeys an exact Taylor expansion:

$$\hat{\alpha} - \alpha_0 = 2\gamma_0' G_n + G_n' G_n = Z_n/\sqrt{n} + \frac{p}{n} + r_n,$$

where $Z_n \sim N(0, 4)$, and $nG_n'G_n \sim \chi^2_p$. Then, $\mathrm{E}G_n'G_n = p/n$, and $r_n = G_n'G_n - p/n$ obeys $\sqrt{\mathrm{E}r_n^2} \lesssim p^{1/2}/n$, so that $r_n = O_P(p^{1/2}/n)$ by Markov inequality. ∎

This example shows that the bias is the bottleneck. If we remove the bias somehow, then we can improve the requirement from $p^2/n \to 0$ to a weaker condition.

There are several ways of removing the bias:

a) *Analytical bias correction*, where we estimate b/n using analytical expressions for the bias and set

$$\breve{\alpha} = \hat{\alpha} - \hat{b}/n.$$

In Example 1 above, the bias corrected estimator is just $\hat{\alpha} - p/n$, but the expressions are much more complex in real problems.

b) *Split-sample bias correction*, where we split the sample into two parts, compute the estimator on the two parts $\hat{\alpha}_{(1)}$ and $\hat{\alpha}_{(2)}$, then set $\bar{\alpha} = \frac{1}{2}\hat{\alpha}_{(1)} + \frac{1}{2}\hat{\alpha}_{(2)}$, and then set

$$\breve{\alpha} = \hat{\alpha} - (\bar{\alpha} - \hat{\alpha}) = 2\hat{\alpha} - \bar{\alpha}.$$

Here we can average over many splits to reduce variability, and it is also possible to use the bootstrap and leave-one-out methods for bias correction.

Why does the sample-splitting method work? The first order biases of $\hat{\alpha}$, $\hat{\alpha}_{(1)}$, and $\hat{\alpha}_{(2)}$ are

$$\frac{b}{n}, \quad \frac{b}{n/2}, \quad \frac{b}{n/2},$$

so that the first order bias of $\breve{\alpha}$ is

$$2\frac{b}{n} - \left(\frac{1}{2}\left[\frac{b}{n/2}\right] + \frac{1}{2}\left[\frac{b}{n/2}\right]\right) = 0.$$

With the bias correction, the resulting side conditions are weaker.

Assertion 2 (Bias-Corrected GMM Inference with Moderately Many Parameters and Moments)**.** *There exist regularity conditions such that if the dimensionality and the number of moments is small compared to the sample size, namely that:*

$$(p+m)^{3/2}/n \to 0 \text{ as } n \to \infty,$$

then the approximate normality and consistency results for the bias-corrected GMM estimator continue to hold:

$$\sqrt{n}(\breve{\alpha} - \alpha) \overset{\text{a}}{\sim} N(0, V_{11}),$$

where V_{11} is the upper-left block of V in (2.2).

References

[1] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

[2] Matias Cattaneo, Michael Jansson, and Whitney Newey. Alternative asymptotics and the partially linear model with many regressors. CeMMAP working papers CWP36/15, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, July 2015.

[3] V. Chernozhukov, D. Chetverikov, and K. Kato. Central Limit Theorems and Bootstrap in High Dimensions. *Annals of Probability*, 2016. forthcoming.

[4] Jinyong Hahn and Guido Kuersteiner. Bias reduction for dynamic nonlinear panel models with fixed effects. *Econometric Theory*, 27(06):1152–1191, 2011.

[5] Jinyong Hahn and Whitney Newey. Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica*, 72(4):1295–1319, 2004.

[6] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.

[7] Whitney K. Newey and Frank Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009.