Method of Moments and Maximum Likelihood in the Laboratory

David K. Levine[1]

Abstract

I argue that specification error is a feature not a bug. In the presence of specification error maximum likelihood never-the-less minimizes the Kullback–Leibler divergence. I argue that this is a poor measure of the distance of a theory from data and that consequently maximum likelihood has poor robustness properties with respect to specification error. I define the weak convergence based notion of specification consistency and show that while maximum likelihood is not generally specification consistent, the method of moments is. The lack of robustness of maximum likelihood is especially problematic with the types of theoretical models used to analyze laboratory data.

[1]Department of Economics, RHUL

Acknowledgements: First version: October 7, 2024. I would like to thank Roberto Corrao, Drew Fudenberg, Michael Mandler, Agnieszka Mensfelt, Tom Palfrey, Kostas Stathis, and Vince Trencsenyi. The Leverhulme Trust provided financial support for which I am grateful. A particular debt is owed to Nikos Nikiforakis and Hans-Theo Normann for providing me with their original experimental data.

> like Copernicus, Lucas thought that a beautiful simple model that fits less well than a more complicated ugly model is somehow closer to the truth [Sargent (2024)]

1. Introduction

In a public goods experiment conducted by Nikiforakis and Normann (2008) four participants are given 20 tokens each worth roughly 7.5 cents. The tokens can be kept or contributed to a common pool in which case their value is multiplied by 0.4 and this number of tokens is awarded to every player. Each group of four participants played ten times and there were a total of six sessions: the empirical distribution of contributions with the 120 observations over the last five rounds of play have a cluster at zero and are reported below in Figure 1.1.

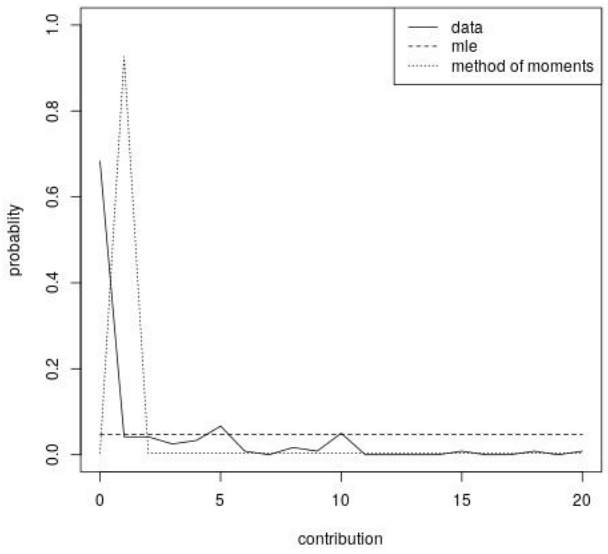

Figure 1.1: Data and Estimates

An investigator would like to know if a model of identical slightly altruistic players each willing to contribute a token, but who have an unknown probability β of trembling uniformly over $\{0, 1, \ldots, 20\}$, does a good job of describing the data. To do this the first step is to estimate β from the data.

Not being a statistician the investigator runs down the hall to the office of R. A. Fisher and asks how to proceed. "Do maximum likelihood" says Fisher. The

likelihood function is graphed in Figure 1.2 below. The maximum likelihood estimate is $\hat{\beta} = 1$ resulting in a prediction that contributions should be uniformly distributed.[2] This uniform distribution is shown above in Figure 1.1. The investigator concludes that the model does an extremely poor job of describing the data.

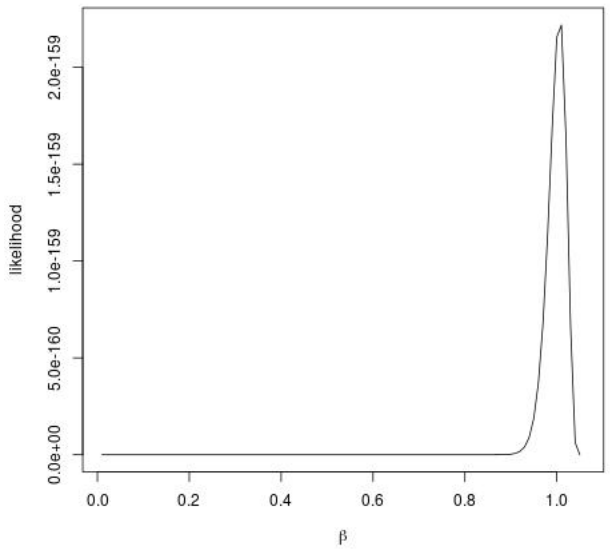

Figure 1.2: Likelihood Function

The investigator has an ambitious research assistant who runs down the hall to the office of K. Pearson[3] who says "Use the method of moments." Matching the expectation generated by the model to the sample average of contributions of 1.7 gives an estimate of $\hat{\beta} = 0.078$ resulting in a spike at 1 shown above in Figure 1.1. The research assistant returns in triumph to the investigator.

I should emphasize that while maximum likelihood does a poor job in this example, any likelihood based approach, including Bayesian inference, has the same difficulty. This can be seen in Figure 1.2 above which plots the likelihood function (not the

[2]The reported maximum likelihood estimate is the constrained estimate $\beta \leq 1$. The likelihood function remains well-defined for larger values of β, however, as the unconstrained estimate is $\hat{\beta} = 1.01$, this makes little difference.

[3]Fisher and Pearson, two of the founders of the discipline of mathematical statistics, had a heated dispute over whether maximum likelihood or the method of moments is the better procedure. See Pearson (1936).

log-likelihood function) so that it is proportional to the Bayesian posterior with a uniform prior. As can be seen all the weight is on unrealistically high values of β.

What is happening in this example? It is highly unlikely that the data is generated by sampling error from a uniform distribution over $\{0, 1, \ldots, 20\}$: the probability of drawing 82 zeroes in 120 observations is about 2×10^{-14}. Rather the issue here is specification error: the model contemplates contributions of 1 while they are in fact 0.

The purpose of this paper is to show that the problem in this example is endemic to likelihood based methods. It is due to the fact that in the presence of specification error the Kullback–Leibler divergence is asymptotically minimized by maximum likelihood. I show that this divergence is a poor measure of the similarity between two distributions. It overemphasizes low probability events about which economists rarely are interested; it ignores important aspects of the data; and it fails the fundamental test that if one probability measure converges weakly to another then the divergence should go to zero. By contrast I show that if divergence is given by a measure of the distance between theoretical and empirical moments then weak convergence is if and only if this divergence goes to zero. I conclude that the distance between theoretical and empirical moments is a good measure of fit, and that the method of moments is consequently more robust to specification error than maximum likelihood.

Before proceeding to the analysis, let me discuss the obvious objection to the example: if the model is misspecified, then surely the solution is to find a better specification? For example, include selfish types as well as mildly altruistic types, or replace mildly altruistic types with selfish types. There are a number of reasons this may not be a good idea.

First, as is commonplace to observe, the purpose of models is not to mimic reality but rather to provide a simplified but none-the-less useful guide to reality. The usual example is that of stylized subway maps which do an abysmal job of reflecting reality but are extremely useful. Even in the hard sciences we do not find, for example, physicists studying the quantum interactions of particles using models that include gravitational forces. I am not sure that the clear conclusion from this understanding is always drawn: specification error is not a bug, it is a feature. A simple model that does a good but not perfect job of reflecting reality is a good model.

Second, there may be good reasons for choosing different specifications or more complex models. However, it does not make sense to choose a different model because

a particular estimation procedure has bad properties. That is, while it might be a good idea to introduce selfish types to better explain the data, it is not a good idea to introduce them simply because maximum likelihood estimation will behave better.

Third, let me indicate that, in the example, introducing adding selfish types or replacing mildly altruistic types with selfish types is not clearly a good idea. In either case both maximum likelihood and the method of moments will do a good job fitting the data: they will do so by concluding that there are no mildly altruistic types. Unfortunately, while this does a good job of explaining the one treatment I have just described, the Nikiforakis and Normann (2008) experiment included four other treatments in which it is possible, at a cost, to punish free-riders. If there are only selfish players who tremble, the prediction is that this will make no difference to expected contribution levels: they should remain low, on the order of 2. In fact, when there is the possibility of punishment, average contributions[4] are much higher: the average is 13.3. This grossly contradicts both a model of only selfish players and a two-type model estimated from the no-punishment data. By contrast, the mild altruism model, either because players are concerned with fairness as in Fehr and Schmidt (1999), or because they understand the need to discourage free-riding as in Levine (2024), can predict much higher contributions when there is the possibility of punishment. In other words, there is good reason for wanting to know that the mild altruism model does not do too badly in the base model without punishment, and maximum likelihood fails to capture how well it does.

Mathematically, I consider a limit in which the truth and the model approach each other. In this setting, under suitable regularity conditions, I show that the method of moments estimator approaches the truth but that maximum likelihood need not.

In most of the laboratory applications I consider, better and better approximations are relevant. However, there is an important case in which it makes sense to consider a sequence of different truths that approach a fixed theoretical model: this is when the theoretical model has a continuous density and the true model is discrete. For example, choices are on a fine grid and this is modelled with a continuous density. In this case the relevant limit is to consider finer and finer grids so that the truth approaches the theory, rather than the other way around. This is often the case outside the laboratory.

[4]This average is also computed over the final five periods. There are 480 observations.

When the truth approaches a fixed theory I am able to give a positive result concerning maximum likelihood. Suppose that the theoretical model has a continuous strictly positive density function and that an identification condition is satisfied. Then in the limit maximum likelihood approaches the truth. This extends existing results about specification error in maximum likelihood that rely on the Kullback-Leibler divergence. However, I also show that the assumption of a continuous strictly positive density function fails in quite ordinary examples.

2. Literature Review

To begin with, it cannot be the case that maximum likelihood always has different properties than the method of moments. In the linear regression model with normally distributed errors maximum likelihood minimizes mean square error, so it is a method of moments estimater and shares the good properties of the method of moments. Moreover, the method of moments is not a cure for severe specification error: for example, as Heckman (1979) points out, sample selection bias can be a severe problem even in the linear regression model.

More broadly, however, maximum likelihood diverges from the method of moments, and there is a large literature examining the impact of specification error on maximum likelihood. Without reviewing all of the many papers written on the subject, the central point is that these papers focus on relative entropy, the Kullback-Leibler divergence, because maximum likelihood asymptotically minimizes this in the presence of specification error.

In time series analysis, robust control theory, as in Hansen and Sargent (2001) and Hansen et al (2006), attempts to find control methods that are robust to perturbations in the model. This theory is based on perturbation that are small as measured by Kullback-Leibler divergence. Watson and Holmes (2017) provide a good overview of this literature.

There is also a literature that asks when maximum likelihood is consistent despite the presence of specification error. An early paper is Gourieroux, Monfort and Trognon (1984), which gives sufficient conditions for MLE to give consistent estimate of moments in a misspecified model. There are two points to be made about this. First, the family studied is the linear exponential family, so is absolutely continuous with respect to the base model: this limits the class of perturbations to those that have finite Kullback-Leibler divergence. I will show that this is very limiting. Second,

if the goal is to get consistent estimates of the moments, the method of moments, rather than maximum likelihood, seems like an obvious approach.

Maximum likelihood with specification error is sometimes called quasi-maximum likelihood or pseudo maximum likelihood. Newey and Steigerwald (1997) give conditions under which quasi-maximum likelihood can be used with non-normal distributions. Again the class of perturbations they consider have a continuous density so finite Kullback-Leibler divergence.

There is also a literature showing that maximum likelihood shares the property of ordinary least squares that estimates remain consistent in the face of a misspecified variance-covariance structure of the errors. This is discussed, for example, in Levine (1983) and Andrews (1991). Again, this departure from the theory is well measured by the Kullback-Leibler divergence.

I should indicate that the literature generally considers a fixed theoretical model and sequence of truths that converge to the theory. By contrast I consider the more general case where both the theory and the truth converge to each other.

3. The Setup

The object of interest is the outcome of the experiment which is taken to be a vector y. This could be the level of contribution of an individual to a public good experiment (a scalar) or it could be a vector representing the play of two players in a match, and so forth. The possible outcomes are bounded and specifically space of possible outcomes Y is taken to be a compact subset of a finite-dimensional vector space. This space is endowed with the usual topology and σ-algebra.

The goal is to model the distribution of outcomes in the underlying population from which participants are drawn in an iid fashion. This distribution F lies in $\mathcal{F}$ the space of probability measures over Y. A distribution $F \in \mathcal{F}$ represents the random draw of a participant. Each choice of that participant can be represented by a random variable, a scalar valued integrable function $g : Y \to \Re$. Denote the expectation of that random variable by $Eg|F \equiv \int g(y)dF(y)$.

Note that in the laboratory there is generally a single geographical location and all observations are collected from the same population at (roughly) the same time. As a result it is rare to condition on exogenous variables x when working with laboratory data. For this reason and to make the exposition more transparent, I will not consider

exogenous variables in the text. Appendix III shows how the results extend to the case of exogenous variables.

Weak Convergence

By way of background, a sequence of distributions *converges weakly,* $F^n \to F$, if for every bounded continuous function $g : y \to \Re$ it is the case that $Eg|F^n \to Eg|F$. It is important here that g is continuous: this means that it preserves the notion of closeness in the underlying space X.

Weak convergence is the gold standard of convergence for probability measures. Three examples illustrate this. First, weak convergence preserves ordinary convergence. That is, if $y^n \to y$, then a point mass on y^n converges weakly to a point mass on y. Second, if F^n is a discrete uniform distribution with support on n points then F^n converges weakly to the continuous uniform distribution. Finally, if you think that an average is approximately normally distributed in a large sample then you agree that weak convergence is the gold standard. Specifically, if $\tilde{y}^n$ are random variables with zero mean and unit variance then the distribution of $(1/T)\sum_{t=1}^{T} \tilde{y}^n$ converges weakly to the standard normal.

Because Y is compact the space of probability distributions $\mathcal{F}$ over Y is compact with respect to the topology induced by weak convergence. This means that sequences F^n of probability distributions have subsequences that converge weakly to a limiting distribution.

The classical reference on weak convergence of probability measures is the book by Billingsley (1968).

Total Variation

There are other notions of the convergence of probability measures, the most important of which is the total variation. The distance between two probability measures in *total variation* is defined as $\left\|F - \tilde{F}\right\| = \sup_A |F(A) - \tilde{F}(A)|$ where the sup is taken over all Borel subsets of Y. However, convergence of total variation fails in all three of the examples above. If $y \neq z$ then $\|y - z\| = 1$; if F^n is a discrete uniform and F the continuous uniform then $\|F^n - F\| = 1$; and if $\tilde{y}^n$ are normalized binomial averages and F standard normal then $\|F^n - F\| = 1$.

From the Portmanteau Theorem if $\|F^n - F\| \to 0$ then also F^n converges weakly to F.

Kullback-Leibler Divergence

The next step is to assess how maximum likelihood behaves in a misspecified model. By way of background, the measure F is absolutely continuous with respect to F_0 if $F_0(A) = 0$ implies $F(A) = 0$. The Radon–Nikodym theorem says that in this case there is a non-negative measurable density function $f : y \to \Re$ such that $F(A) = \int_A f(y)dF_0(y)$

Suppose that $F, \tilde{F}$ are absolutely continuous with respect to a common F_0 so have densities $f, \tilde{f}$. If $\tilde{F}$ is absolutely continuous with respect to F then the *Kullback–Leibler divergence* is defined by $D_{KL}(\tilde{F}|F) = \int \tilde{f}(y) \log(\tilde{f}(y)/f(y))dF_0(y)$ otherwise $D_{KL}(\tilde{F}|F) \equiv \infty$. This is non-negative and is zero if and only if $F = \tilde{F}$, but it is not a metric because it is not symmetric. For completeness I prove this in Proposition 7.1 in Appendix I. From Pinsker's Inequality if $D_{KL}(F^n|F) \to 0$ or $D_{KL}(F|F^n) \to 0$ then $\|F^n - F\| \to 0$ and this in turn implies F^n converges weakly to F.

Often F_0 is the uniform distribution, either continuous or discrete, so that f and $\tilde{f}$ are the ordinary continuous or discrete density functions. However the theory can accomodate mixed continuous discrete distributions such as the Tobit. The Kullback–Leibler divergence is an essential tool for understanding the behavior of maximum likelihood in a misspecified model.

If $F(\beta)$ is a parametrized family of distributions and samples are drawn from a distribution $\tilde{F}$ then asymptotically maximum likelihood minimizes the Kullback–Leibler divergence $D_{KL}(\tilde{F}|F(\beta))$. Although it is more generally true, it is easy to see in the case where $\tilde{F}$ is discrete, that is, observations lie in a finite set Y. In this case the log-likelihood in the sample is proportional to $\sum_{y \in Y} \phi(y) \log f(y|\beta)$ where $\phi(y)$ is the fraction of observations on y in the sample. Asymptotically this converges in probability to

$$\sum_{y \in Y} \tilde{f}(y) \log f(y|\beta) = -D_{KL}(\tilde{F}|F(\beta)) + \sum_{y \in Y} \tilde{f}(y) \log \tilde{f}(y)$$

Convergence is uniform in β since Y is finite and f bounded. Hence any sequence of maximizers of the log-likelihood function has a limit point that is a minimizer of the Kullback–Leibler divergence.

4. An Assessment of the Kullback–Leibler Divergence

I now examine whether the Kullback-Leibler divergence is a useful measure of the distance between two distributions. In discussing weak convergence versus total variation I showed that there are three simple tests which weak convergence passes and total variation fails. Like total variation, Kullback-Leibler divergence fails all three tests.

Specifically, if $y \neq z$ then $D_{KL}(y|z) = \infty$; if F^n is a discrete uniform and F the continuous uniform then $D_{KL}(F^n|F) = D_{KL}(F|F^n) = \infty$; and if $\tilde{y}^n$ are normalized binomial averages and F is standard normal then $D_{KL}(F^n|F) = D_{KL}(F|F^n) = \infty$. All of these results are special cases of mutually singular distributions.

Two distributions $F, \tilde{F}$ are *mutually singular* if there are two disjoint sets $A \cup B = Y$ such that $F(A) = \tilde{F}(B) = 0$. In each of the test cases F^n and F are mutually singular: take A to be the support of F^n.

It is apparent if $F, \tilde{F}$ are mutually singular $\left\| F - \tilde{F} \right\| = 1$. In addition since $F(B) = \tilde{F}(A) = 1$ neither is absolutely continuous with respect to the other, so $D_{KL}(\tilde{F}|F) = D_{KL}(F|\tilde{F}) = \infty$.

Continuous distributions can be approximated by discrete distributions in the sense of weak convergence. However, a discrete and continuous distribution are mutually singular, so the Kullback–Leibler divergence provides no useful information about this convergence.

Labels Matter

Example 4.1. Consider a simple variation on the motivating example in the introduction. Take Y to be the evenly spaced n-grid on $[0, 20]$ where $n \geq 21$ and $\beta_0 \in [0, 1]$. Let F^n be the "warm glow of giving" model that places weight $1 - \beta + \beta/n$ on the first strictly positive grid point and β/n on the remainder. Suppose that the data is given by $\tilde{F}^n(\beta_0)$ which has a mass point at 0 with probability $1 - \beta_0 + \beta_0/n$ with the remaining points each having probability β_0/n.

The expected log-likelihood function for this model is given by $(1-\beta_0/n)\log(\beta/n) + (\beta_0/n)\log(1 - \beta + \beta/n)$. The derivative with respect to β is

$$\frac{(1/n)(1 - \beta_0/n)}{\beta/n} - \frac{(1 - (1/n))(\beta_0/n)}{1 - \beta + \beta/n},$$

and the second derivative is negative so that the maximum is characterized by the first

order condition. Evaluating the derivative at $\beta = 1$ gives $1-\beta_0$. Hence the constrained asymptotic maximum likelihood estimator is $\hat{\beta} = 1$, and the unconstrained estimator bigger than 1.

The method of moments equates the actual mean of $10\beta_0$ to the theoretical mean $(1-\beta)(21/n) + 10\beta$ giving

$$\hat{\beta} = \beta_0 - (21/n)\frac{1-\beta_0}{10-21/n},$$

slightly less than the true value. (In the example $\hat{\beta}$ is slightly higher than the empirical value due to the fact that the data is not in fact uniform.) Here a slightly too low spike is predicted at a slightly too high contribution level.

So far this analysis has added nothing to the example in the introduction. The point here is a different one. Consider an alternative theoretical model, a model of high altruism. With high altruism if there is no trembling, that is with $1-\beta$ probability, the participant contributes not $21/n$, but rather 20. The method of moments now yields $\hat{\beta} = 1$ and unless β_0 is large the true mean of $10\beta_0$ is poorly matched by the theoretical mean of 10. Maximum likelihood agrees with this assessment and therein lies the problem.

The likelihood function and the Kullback–Leibler divergence here are exactly the same regardless of whether the theoretical mass point is at $n/21$ or 20. This is despite the fact that the two contribution levels are very different, and despite the fact that the warm glow theory matches the data reasonably well and the high altruism theory does not. While values of y have economic meaning - a contribution of $21/n$ is much closer to 0 than a contribution of 20 - entropy based measures such as the Kullback–Leibler divergence make no use of this information. The labels do not matter to entropy - but for economic analysis they should. According to the Kullback–Leibler divergence the warm glow and high altruism models are equally bad because there is no difference between a contribution of $21/n$ and a contribution of 10. From an economic perspective this is not true. By contrast the method of moments says that the warm glow theory does much better than the high altruism theory - and indeed this is the case.

Sharp Theories

An important theory is subgame perfection with selfish money maximizing players. This is a sharp theory in the sense that it makes definitive predictions, and derives

these predictions without estimation - "out-of-sample" - by reading the experimental instructions. By contrast, level-0 theory (see Stahl and Wilson (1994)) says that players randomize uniformly over all actions. This is not a sharp theory, indeed, a rather fuzzy one, but is also entirely "out-of-sample."

There are three classical experiments in which the general consensus is that subgame perfection does well. These are the best-shot game, the market auction game, and the one-shot prisoner's dilemma game. The best-shot game is a two player sequential public goods contribution game where contributions are in $\{0, 1, \ldots, 21\}$ and only the highest contribution matters. The prediction of subgame perfection is that the first mover never contributes. In the Prasnikar and Roth (1992) there are ten rounds. From the 7th round on, in treatments where participants are fully informed of the extensive form, the first mover indeed never contributes. By contrast the level-0 theory predicts that the first mover will contribute zero only 4.5% of the time.

The market auction game is a ten player game in which nine players submit bids on an object worth \$10.00 in increments of 5 cents. The tenth player can accept or reject the highest bid. The prediction of subgame perfection is that the highest bid will be either \$9.95 or \$10.00 and that this will be accepted. In Roth et al (1991) there are ten rounds. In the final six rounds this prediction is always correct. By contrast, the level-0 theory predicts that the high bid will be \$9.95 or \$10.00 only 8.6% of the time.

In the one-shot prisoner's dilemma game the prediction of subgame perfection is that all players will defect. In the final of ten rounds played against strangers Dal Bo (2005) finds that in fact 94.2% of them do. By contrast, the level-0 theory predictions that only 50% of them should defect.

As indicated: the consensus is that subgame perfection does well in describing participants play in these experiments, and clearly level-0 does not. However: because it is a sharp theory, subgame perfection also predicts that certain things will not happen that in fact do. In best shot it predicts that the second mover will always contribute 4. In fact the average contribution of the second player in the tenth of ten rounds is 3.88. In the market auction subgame perfection predicts that if the high bid is \$9.95 all bids should be \$9.95. This is never the case. In the one-shot prisoner's dilemma subgame perfection predicts no cooperation, while in fact there is a 5.8% cooperation rate.

In the discrete case if for some y the theoretical density value $f(y)$ is zero and the

true density value $\tilde{f}(y) > 0$, no matter how small, then $D_{KL}(\tilde{F}|F) = \infty$. In particular, for all three experiments, the Kullback–Leibler divergence of subgame perfection from the data is infinite. By contrast the level-0 theory predicts everything has positive probability so has a finite Kullback–Leibler divergence. By the logic of maximum likelihood the level-0 theory is the better theory.

Clearly if we modified the theory of subgame perfection to include noise, for example by using a quantal response model as in McKelvey and Palfrey (1995), or a probability of uniform trembling as in the example in the introduction, the resulting theory would better fit the data. However: there there are many choices of how to do this, the resulting theory becomes complicated and not sharp, and the simple theory is already working well. Certainly the fact that maximum likelihood does not work well is not a good reason for complicating a simple and sharp model.

Small Probabilities Should Not Matter

Zero probabilities are special, and if Y is finite we can always add some noise to a theoretical model so that the theoretical probabilities are not zero. However: the problem with zero probabilities is also a problem for very small probabilities.

Example 4.2. Consider the (hypothetical) example below where participants choose between three alternatives A, B, C. In this experiment 90% of participants choose A, 9% choose B and 1% choose C. There are two theories labeled F_1 and F_2. The first theory predicts that 90% of people will choose A and that almost all of the rest will choose B with only a tiny, but non-zero, fraction 10^{-100} choosing C. This is a pretty good theory. By contrast the second theory gets A and B hopeless muddled, with 90% choosing B rather than A and 9% choosing A rather than B - the direct opposite of reality. However: the second theory correctly predicts that 1% of people will choose C.

In Table 4.1 below I report these numbers and in the final row I compute the Kullback–Leibler divergences between the true model and the theories. The point is that the second theory has smaller Kullback–Leibler divergence than the first so is preferred by maximum likelihood. In other words: the entropy based measure cares very little about getting the high probability events A, B correct, and very much about getting the low probability event C right. This stands common sense on its head.

	$\tilde{F}$	F_1	F_2
A	0.90	0.90	0.09
B	0.09	$0.10 - 10^{-100}$	0.90
C	0.01	10^{-100}	.01
$D_{KL}(\tilde{F}\|F)$		2.25	1.87

Table 4.1: KL Divergence

5. Specification Consistency and the Generalized Method of Moments

This paper is about specification error not sampling error, and in the intended applications it should be possible to choose large samples if these are required. For these reasons, I will abstract from sampling error by assuming an infinite sample. More concretely, I will in this section replace sample moments with the corresponding expectations. Appendix II discussed sampling error.

Let $\beta \in B$ a finite dimensional compact parameter space. A sequence of models $F^n(\beta)$, each continuous in β, *approximates* $\tilde{F}^n$ at β_0 if, for all $\beta^n \to \beta$, the approximations $F^n(\beta^n) \to F(\beta)$ and $\tilde{F}^n \to F(\beta_0)$. This implies that the signed measure $F^n(\beta_0) - \tilde{F}^n$ converges weakly to zero. Here $\tilde{F}^n$ may represent different experiments, for example, refining a grid over which choices are made. Note that since $F^n(\beta)$ is continuous and converges uniformly to $F(\beta)$ it follows that $F(\beta)$ is also continuous in β.

If $F^n(\beta)$ approximates F^n at β_0 an *asymptotic estimation procedure* is a sequence of subsets B^n of the parameter space. An asymptotic estimation procedure is *specification consistent* if for any $\hat{\beta}^n \in B^n$ the sequence $F^n(\hat{\beta}^n)$ converges weakly to $F(\beta_0)$. The idea is that B^n corresponds to the limits of estimators in a large sample. Since $F^n(\hat{\beta}^n) \to F(\beta_0)$ and $\tilde{F}^n \to F(\beta_0)$ it follows that the signed measure $F^n(\hat{\beta}^n) - \tilde{F}^n$ converges weakly to zero. Intuitively what this says is if the approximation is good enough then the estimates are approximately correct.

As an example of an asymptotic estimation procedure, suppose that $F^n(\beta)$ has a continuous density function $f^n(y|\beta)$ with respect to some underlying distribution F_0^n. In this case maximum likelihood in which B^n is defined as the maximizers of $E \log f^n(y|\beta)|\tilde{F}^n$ is an asymptotic estimation procedure.

The next two examples show that maximum likelihood is not generally specification consistent. The first example is derived from example 4.1 and the limiting model F has a mass point at zero. However, mass points are not needed, and the second

example uses density functions that are strictly positive and continuous.

Labels Again

Example 5.1. As in 4.1 take Y to be the evenly spaced n-grid on $[0, 20]$ where $n \geq 21$ and take $B = [0, 1]$. As in that example F^n is a "warm glow of giving" model that places weight $1 - \beta + \beta/n$ on the first strictly positive grid point and β/n on other grid points. The "true" model is $\tilde{F}^n$ which has a mass point at 0 with probability $1 - \beta_0 + \beta_0/n$ with the remaining points each having probability β_0/n. Note that this is continuous in β as required.

Define $F(\beta)$ to be a mass point at 0 with probability $1-\beta$ and with the remaining weight of β uniformly distributed over $[0, 20]$. If $\beta^n \to \beta$ then $F^n(\beta^n) \to F(\beta)$. In addition $\tilde{F}^n \to F(\beta_0)$ so that F^n approximates $\tilde{F}^n$ at β_0. As shown previously, asymptotic maximum likelihood when the true distribution is $\tilde{F}^n$ always minimizes $D_{KL}(\tilde{F}^n | F^n(\beta))$ by choosing $\hat{\beta} = 1$ regardless of the true value of β_0. It follows that $F^n(\hat{\beta}^n)$ converges weakly to the uniform distribution on $[0, 20]$ so maximum likelihood is not specification consistent for $\beta_0 < 1$.

Continuous Densities

Example 5.2. Suppose that $Y = [0, 1]$ and that $B = [0, 1]$. Take $\tilde{F} = F(\beta_0)$ and the distributions $F(\beta)$ and $F^n(\beta)$ are given by strictly positive density functions with respect to Lesbesgue measure $f(y|\beta)$, $f^n(y|\beta)$ jointly continuous in (y, β) as described next.

Assume that $n \geq 2$. Specifically, the density $f(y|\beta)$ is linear with $f(1|\beta) = (1 + 99\beta)/100$ and $f(0|\beta) = 2 - f(1|\beta)$. The density $f^n(y|\beta)$ is continuous and piecewise linear with knots at $\beta \in \{1/n, 2/n\}$ with $f^n(0|\beta) = f(1/n|\beta) = e^{-\beta n^2}$, $f^n(1|\beta) = f(1|\beta)$ and

$$f^n(2/n|\beta) = \frac{2 - (3/n)e^{-\beta n^2} - (1 - 2/n)f(1|\beta)}{1 - 1/n}.$$

For $y > 0$ suppose that $\beta^n \to \beta$. Then the densities $\bar{f}^n(y) \equiv f^n(y|\beta^n) \to f(y|\beta)$ converge pointwise and from Lesbesgue's dominated convergence theorem this implies weak convergence of the distributions $F^n(\beta^n) \to F(\beta)$.

The expected log-likelihood function is $L^n(\beta) = \int f(y|\beta_0) \log f^n(y|\beta) dy$. Hence

$L^n(0) \geq (1/100)\log(49/50)$. For any $\beta > 0$

$$\int_{1/n}^{1} f(y|\beta_0)\log f^n(y|\beta)dy \leq 2\log 4.$$

However $\int_0^{1/n} f(y|\beta_0)\log f^n(y|\beta)dy = -(2 - f(1|\beta_0))\,\beta n$, so $L^n(\beta) \leq 2\log 4 - \beta n$. Hence for $n > (1/\beta)(2\log 4 - (1/100)\log(49/50))$ it follows that $\hat{\beta}^n < \beta$, and in particular the asymptotic maximum likelihood estimator $\hat{\beta}^n \to 0$ regardless of β_0, so maximum likelihood is not specification consistent. In particular, $F^n(\hat{\beta}^n)$ converges weakly to a $F(0)$, the linear density $\hat{f}(0) = 199/100, \hat{f}(1) = 1/100$ even if, in fact, $\beta = 1$ and the true density is uniform.

The key point here is that, while convergence of the densities implies weak convergence, the convergence is pointwise but not uniform. If the convergence was uniform, this would imply specification consistency of maximum likelihood. This example is an elaboration of example 4.2 involving small probabilities. Here, on the left of the theoretical model, the probabilities are tiny and this forces maximum likelihood to make poor choices.

The Method of Moments is Specification Consistent

Let $\mu^n(y|\beta)$ be ℓ-dimensional vector valued functions jointly continuous in (y, β). Denote by Γ the convex hull of $\mu(Y)$. Let $h : \Gamma \to \Re_+$ be non-negative, continuous, and satisfy $h(\gamma) = 0$ if and only if $\gamma = 0$. One obvious example is $h(\gamma) = \|\gamma\|^2$ for some norm on $\Re^\ell$, but other functions can be used. The *generalized method of moments* with respect to μ^n, h is the asymptotic estimation procedure given by

$$B^n = \arg\min_{\beta} h\left(E\mu^n(\beta)\middle|\tilde{F}^n\right).$$

The *ordinary method of moments* is defined by $\nu(y)$, an ℓ-dimensional vector valued function with $\mu^n(y|\beta) = \nu(y) - E\nu|F^n(\beta)$. Since $F^n(\beta)$ is assumed to be continuous in β, $F^n(\beta^t) \to F^n(\beta)$ so, as ν is continuous, $E\nu|F^n(\beta^t) \to E\nu|F^n(\beta)$. Consequently $\mu^n(y|\beta)$ is jointly continuous in (y, β), thus the ordinary method of moments is a special case of the generalized method of moments.

Call μ^n *convergent* if it converges uniformly to μ, and $E\mu(\beta_0)|F(\beta_0) = 0$. From the uniform limit theorem since the μ^n are continuous so is μ.

Proposition 5.3. *The ordinary method of moments is convergent.*

Proof. By definition $\mu^n(y|\beta) = \nu(y) - E\nu|F^n(\beta)$. Define $\mu(y|\beta) \equiv \nu(y) - E\nu|F(\beta)$. Then ν converges uniformly to itself and since for any $\beta^n \to \beta$ by assumption $F^n(\beta^n) \to F(\beta)$ it follows that $E\nu|F^n(\beta^n) \to E\nu|F(\beta)$, that is the convergence of $F^n(\beta)$ to $F(\beta)$ forces the uniform convergence of $E\nu|F^n(\beta)$ to $E\nu|F(\beta)$. Finally, $E\mu(\beta_0)|F(\beta_0) = E\nu(y)|F(\beta_0) - E\nu|F(\beta_0) = 0$. □

Say that $F(\beta)$ is *identified* with respect to μ if, for any $\beta \in B$, $E\mu(\beta)|F(\beta_0) = 0$ implies $\beta = \beta_0$. To illustrate, consider the ordinary method of moments with $\nu(y) = y$. In the first example $E\mu(\beta)|F(\beta_0) = 10(\beta - \beta_0)$ so that model is identified with respect to μ. In the second example $E\mu(\beta)|F(\beta_0) = (201 + 99(\beta - \beta_0))/600$ so that model is also identified with respect to μ.

Note that if a model is identified then specification consistency is if and only if for all $\hat{\beta}^n \in B^n$ it is the case that $\hat{\beta}^n \to \beta_0$.

Theorem 5.4. *Suppose that $F^n(\beta)$ approximates $\tilde{F}^n$ at β_0, that μ^n is convergent, and that $F(\beta)$ is identified with respect to μ. Then the generalized method of moments with respect to μ^n, h is specification consistent.*

Proof. It suffices to show that $\hat{\beta}^n \to \beta_0$. Because B is compact it necessary only to show that if $\hat{\beta}^n$ converges, it converges to β_0.

Suppose that $\beta^n \to \beta$. Then

$$E\mu^n(\beta^n)|\tilde{F}^n = E\left(\mu^n(\beta^n) - \mu(\beta)\right)|\tilde{F}^n + E\mu(\beta)|\tilde{F}^n.$$

The first term converges to zero since μ^n converges uniformly to μ and the second to $E\mu(\beta)|F(\beta_0)$ because $\mu(y|\beta)$ is continuous in y and $\tilde{F}^n$ converges weakly to $F(\beta_0)$. Hence

$$E\mu^n(\hat{\beta}^n)|\tilde{F}^n \to E\mu(\hat{\beta})|F(\beta_0). \tag{5.1}$$

Next, using the fact that the sequence is approximating, it must be that $F^n(\beta_0) \to F(\beta_0)$ and $\tilde{F}^n \to F(\beta_0)$. From equation 5.1 with $\beta^n = \beta_0$ if follows that $E\mu^n(\beta_0)|\tilde{F}^n \to E\mu(\beta_0)|F(\beta_0)$. Since weak convergence implies convergence of the moments and h is continuous with $h(0) = 0$

$$\lim h\left(E\mu^n(\beta_0)|\tilde{F}^n\right) = h\left(E\mu(\beta_0)|F(\beta_0)\right) = 0.$$

Since

$$h\left(E\mu^n(\hat{\beta}^n)\backslash\tilde{F}^n\right) \leq h\left(E\mu^n(\beta_0)\backslash\tilde{F}^n\right)$$

and h is non-negative and continuous

$$h\left(\lim E\mu^n(\hat{\beta}^n)\backslash\tilde{F}^n\right) = \lim h\left(E\mu^n(\hat{\beta}^n)\backslash\tilde{F}^n\right) \leq \lim h\left(E\mu^n(\beta_0)\backslash\tilde{F}^n\right) = 0.$$

Applying equation 5.1 to $\beta^n = \hat{\beta}^n$ this implies that $h(E\mu(\hat{\beta})\backslash F(\beta_0)) = 0$.

From the last equation and by the definition of h if follows that $E\mu(\hat{\beta})\backslash F(\beta_0) = 0$. The identification condition then implies that $\hat{\beta} = \beta_0$. Hence $\hat{\beta}^n \to \beta_0$ implying specification consistency. □

Appendix II shows that Theorem 5.4 remains true in a large enough sample when the asymptotic limits $\hat{\beta}^n$ are replaced with estimates derived from sample averages.

A Remark on Specification

The theoretical model is specified as $F^n(\beta)$ giving rise to the moments $\overline{\mu}^n(\beta) \equiv E\mu(\beta)\backslash F^n(\beta)$ as a function of β. It should be clear that any other $\hat{F}^n(\beta)$ satisfying $E\mu(\beta)\backslash\hat{F}^n(\beta) = \overline{\mu}^n(\beta)$ will give the same results as $F^n(\beta)$, and in particular the generalized method of moments requires specifying only $\overline{\mu}^n(\beta)$ and so is robust with respect to any specification error, large or small, that leaves $\overline{\mu}^n(\beta)$ intact.

Measurement versus Estimation

There are many ways of choosing the particular moments μ that will be used in the generalized method of moments estimation. I view this as a feature not a bug: it enables the investigator to set priorities for which moments are economically important, and it enables the reader to see whether the investigator has cherry-picked obscure moments.

Moments are important not only for estimation, but also for measurement. That is, whatever estimation technique is employed, reporting the theoretical versus the sample moments provides a good way of assessing how successful the model is. For example, in studying self-confirming equilibrium expected losses can be report as suggested in Fudenberg and Levine (1997). In studying behavior mechanism design expected welfare can be reported as suggested by Levine (2024). In studying the

repeated prisoner's dilemma the fraction of participants who cooperate can be reported and has been used by investigators such as Dal Bo (2005) and Fudenberg and Karreskog Rehbinder (2024).

Measurement is particularly important because ideally theories are not estimated but make predictions based on the experimental instructions: subgame perfect Nash equilibrium is such a theory, as are the theories in Levine (1986), Fehr and Schmidt (1999) and Levine (2024). The point is that moments, not likelihoods or divergences, are a good measure of success.

6. Robustness of Maximum Likelihood

If $F^n(\beta)$ has a density function that is sufficiently regular then maximum likelihood is a generalized method of moments estimator. Specifically, say that $F^n(\beta)$ is *continuous,* if there is a density function $f^n(y|\beta)$ with respect to some F_0^n, that is $F^n(A|\beta) = \int_A f^n(y|\beta)dF_0^n(y)$, and $f^n(y|\beta)$ is strictly positive and jointly continuous in (y,β). Then $\mu^n(y|\beta) \equiv \max_\beta E\log(f^n(y|\beta)|\tilde{F}^n - \log(f^n(y|\beta)$ is jointly continuous in (y,β). Minimizing $h\left(E\mu^n(\beta)|\tilde{F}^n\right)$ with respect to β maximizes the likelihood function, so in this case maximum likelihood is a generalized method of moments estimator.

Examples 5.1 and 5.2 show the limitations of this approach: in neither case is μ^n convergent. In example 5.1 $F^n(\beta)$ is an n-grid in $[0,1]$ and F_0^n can be taken to have positive mass on the grid points and zero everywhere else. The density is then defined in the ordinary way at the grid points, and can be extended by linear extrapolation between the grid points. So far so good: but $f^n(y|\beta)$ converges to a mass with probability $1-\beta$ at 0 and a continuous uniform elsewhere and this is not continuous so the convergence is not uniform. In example 5.2, as shown, the densities also converge but not uniformly. While uniform weak convergence of $F^n(\beta)$ to $F(\beta)$ implies that the ordinary method of moments is convergent it does not imply that maximum likelihood is convergent.

The primary conceptual experiment has been that of a fixed truth and better approximations. The notion of approximation is general enough to allow another conceptual experiment, and one that is one widely used. That is to fix the model and consider truths that better approximate the model. This makes sense, for example, if the model has a continuous density function and the truths are discrete over finer

and finer grids. In this case, with $F^n = F$, maximum likelihood is better behaved since certainly μ^n converges uniformly to itself.

Say that $F(\beta)$ is *identified* if for any pair $\beta, \tilde{\beta} \in B$ it is the case that $F(\beta) = F(\tilde{\beta})$ implies $\beta = \tilde{\beta}$.

Theorem 6.1. *Suppose that $F^n(\beta) = F(\beta)$ approximates $\tilde{F}^n$ at β_0 and that $F(\beta)$ is identified and continuous. Then maximum likelihood is specification consistent.*

Proof. Define $\mu(y|\beta) \equiv \log f(y|\beta) - \max_\beta E \log f(y|\beta)$. Theorem 5.4 then applies provided that $E\mu(\beta_0)\text{¦}F(\beta_0) = 0$ and for any $\beta \in B$, $E\mu(\beta)\text{¦}F(\beta_0) = 0$ implies $\beta = \beta_0$. That is to say: $E \log f(y|\beta_0)\text{¦}F(\beta_0) \geq E \log f(y|\beta)\text{¦}F(\beta_0)$ with equality only if $F(\beta) = F(\beta_0)$. This follows from $D_{KL}(F(\beta_0)|F(\beta)) = E \log f(y|\beta_0)\text{¦}F(\beta_0) - E \log f(y|\beta)\text{¦}F(\beta_0)$ and Proposition 7.1 in Appendix I. □

Notice that there is an asymmetry. In example 5.2 $D_{KL}(F(\beta_0)|F^n(\beta_0)) \to \infty$ and maximum likelihood is not specification consistent. By contrast if F is the theoretical model and $F^n(\beta_0)$ generates the data $D_{KL}(F^n(\beta_0)|F(\beta_0)) \to 0$ and indeed maximum likelihood is specification consistent by Theorem 6.1.

The existence of a strictly positive continuous density function $f(y|\beta)$ with respect to some F_0 is important and not always satisfied, as the next three simple examples show. In the first, there is no density function. In the second, there is a strictly positive density function, but it is not continuous. In the third, there is a density function that vanishes at a single point. In all cases the conclusion of Theorem 6.1 fails so that maximum likelihood is not robust, yet Theorem 5.4 implies that an appropriate ordinary method of moments is specification consistent.

No Density

Example 6.2. Consider a continuous version of example 4.1 where the degree of altruism is unknown. Specifically, suppose that $Y = [0, 20]$, $B = [0, 1] \times [0, 20]$ and that the theoretical model $F(\beta)$ has a spike at β_2 with probability $1 - \beta_1$ and trembles uniformly over $[0, 20]$ with probability β_1. The data is generated by $\tilde{F}^n$ on an n-grid in $[0, 21]$. With some probability the two grid points adjacent to β_0 are chosen and with the remaining probability the remaining grid points are chosen such that in the limit as $n \to \infty$ the probability distribution over those points weakly converges to the uniform distribution over $[0, 20]$. Hence $\tilde{F}^n$ converges weakly to $F(\beta_0)$.

Because $F(\beta)$ has a spike at β_2 to be absolutely continuous with respect to a measure F_0 it must be for $\beta_1 < 1$ that $F_0(\beta_1, \beta_2) > 0$. However, there is no single measure F_0 for which this is true for all $\beta_2 \in [0, 20]$. Hence there is no density function, hence no likelihood function, so maximum likelihood estimation is impossible. In general this will be the case when the theoretical model has an atom (or atoms) which move continuously with the parameters.

Note that the model F is not identified with respect to the mean only, but is identified with respect to the mean and variance. Hence if a these two moments are used together with an appropriate h the assumptions of Theorem 5.4 are satisfied and the ordinary method of moments is specification consistent.

Discontinuous Density

Example 6.3. Now reverse example 4.1, so that the theory is that participants are selfish, while in fact they are mildly altruistic. Specifically, $Y = [0, 20]$ and $B = [1/10, 9/10]$. Here $F(\beta)$ is a "selfish" model that places weight $1 - \beta$ on 0 and with probability β is uniform. The true model is $\tilde{F}^n$ which has a mass point at $1/n$ with probability $1 - \beta_0$ and with probability β_0 is uniform. Hence the true model converges weakly and uniformly to the theoretical model.

The theoretical model $F(\beta)$ has a density with respect to the measure F_0 which places weight $1/2$ on 0 and with probability $1/2$ is uniform. The strictly positive density is given by $f(0|\beta) = 2(1 - \beta)$ and for $y > 0$ by $f(y|\beta) = \beta/10$. Hence the expected likelihood function for each n is given by $\log(\beta/10)$, which is maximized at $\hat{\beta} = 9/10$ and the true model does not converge to this for $\beta_0 < 9/10$. However, $Ey = 10\beta$ so the ordinary method of moments with respect to $\nu(y) = y$ is identified, so Theorem 5.4 implies that it is specification consistent

The problem here is that the density function is not continuous. That is, $2(1-\beta) = f(0|\beta) = \lim_{y \to 0^+} f(y|\beta) = \beta/10$ only for $\beta = 20/21 > 9/10$.

Zero Density

Example 6.4. Suppose that $Y = [0, 1]$ and $B = [0, 1]$. The theoretical model has the density $f(y|\beta) = (\beta + 1)y^\beta$ while the data is generated by $\tilde{F}^n$ which is zero with probability $1/n$ and drawn from $f(y|\beta_0)$ with probability $(n - 1)/n$. Then the expected log-likelihood is $-\infty$ regardless of β so $B^n = B$ and maximum likelihood is not robust. Again, as $Ey = (\beta + 1)/(\beta + 2)$ which is strictly increasing in β, this model is not problematic for the ordinary method of moments with $\nu(y) = y$.

7. Conclusion

Simple, sharp models are good models. In the laboratory participants often cluster on a small number of choices, as in the Nikiforakis and Normann (2008) data described in the introduction. Consequently, attention has focused on theoretical models with a small number of types. Subgame perfection and many behavioral models often assume a single type. Other models have several, but still a small number of types: Levine (1986) has three types, Fehr and Schmidt (1999) twelve, and Levine (2024) three. The level-k reasoning model of Stahl and Wilson (1994) is similarly used with a small number of types. These are examples of simple sharp models.

Maximum likelihood has good efficiency properties when there is no specification error. However, simple, sharp models necessarily have specification error. Unfortunately this means that maximum likelihood can provide misleading results. By contrast, I have shown that the method of moments provides robust results for models that only approximate the truth.

The usual argument in favor of maximum likelihood is that if the model is specified correctly it converges (probabilistically) to the true limit at least as fast as any other estimator. This argument loses its force in the universal case of specification error. It may well be preferable to converge less fast to a limit that well approximates the truth rather than very fast to one that does not.

I should conclude by saying that I am not arguing "never do maximum likelihood" or to throw away all the many empirical studies that have used maximum likelihood. Indeed, I focus on experimental data for a reason. Outside the laboratory the most common sort of model is one with a continuous density function, and if we think of that as an approximation to data drawn from an underlying discrete density function then Theorem 6.1 shows that maximum likelihood is specification consistent. In the laboratory, as in the examples, choices are often on grids that can be refined and models with a small number of types are common, and here maximum likelihood is more problematic.

The bottom line is that the method of moments is always a safe choice, and the results here provide a guide to when maximum likelihood is.

Appendix I: Properties of the Kullback–Leibler Divergence

Proposition 7.1. *$D_{KL}(\tilde{F}|F) \geq 0$ and $D_{KL}(\tilde{F}|F) = 0$ implies $F = \tilde{F}$.*

Proof. Recall that $\tilde{F}$ should be absolutely continuous with respect to F or else $D_{KL}(\tilde{F}|F) = \infty$. When absolute continuity is satisfied $D_{KL}(\tilde{F}|F) \equiv -\int \tilde{f}(y) \log(f(y)/\tilde{f}(y)) dF_0(y)$ where by convention the integral is over the region where $\tilde{f}(y) > 0$ and in particular it may be assumed that neither density vanishes in the region of integration.

Define $q(x) = x - 1 - \log x$. This is non-negative and vanishes only at $x = 1$. Since densities must integrate to one, $\int \left(f(y)/\tilde{f}(y) - 1 \right) \tilde{f}(y) dF_0(y) = 0$ so that $D_{KL}(\tilde{F}|F) = \int \tilde{f}(y) q(f(y)/\tilde{f}(y)) dF_0(y)$. This is clearly non-negative, proving the first assertion.

If $D_{KL}(\tilde{F}|F) = 0$, the set Y^+ on which $q(f(y)/\tilde{f}(y)) > 0$ must have $\tilde{F}$ measure zero. Since q vanishes only at 1 this means that $f(y) = \tilde{f}(y)$ on $Y - Y^+$. Then

$$\int_{Y=Y^+} f(y) dF_0 = \int_{Y=Y^+} \tilde{f}(y) dF_0 = 1$$

so Y^+ also has F measure zero. Hence $\tilde{F} = F$. □

Appendix II: Sampling Error and the Generalized Method of Moments

In a finite sample it is necessary to replace the theoretical moment $E\mu^n(\beta)|\tilde{F}^n$ with the sample average $\mu^{nT}(\beta)$ drawn from $\tilde{F}^n$. It is well-known that under suitable regularity conditions this converges uniformly to the theoretical moment. In the current context this implies

Theorem 7.2. *Suppose that $F^n(\beta)$ approximates $\tilde{F}^n$ at β_0, that μ^n, h is a generalized methods of moments estimator, that μ^n is convergent, and that $F(\beta)$ is identified with respect to μ. Then there exists a $\overline{T}_n$ such that for $T_n \geq \overline{T}_n$ any $\hat{\beta}^{nT_n} \in \arg\min_\beta h(\mu^{nT}(\beta))$ satisfies $\hat{\beta}^{nT_n} \to \beta_o$ in probability.*

Proof. The proof has two steps. The first is the well-known result that for fixed n as $T \to \infty$ the limit points of $\hat{\beta}^{nT}$ almost surely lie in B^n. This implies also convergence in probability. Supposing this is the case, the main result then follows from a standard diagonalization argument. Fix some ϵ. Then for $T_n \geq \overline{T}_n$ with probability at least $1 - \epsilon$ it must be that $\hat{\beta}^{nT_n}$ is within $\epsilon/2$ of B^n. Then letting $n \to \infty$ it follows from Theorem 5.4 that $\hat{\beta}^{nT_n}$ with probability at least $1 - \epsilon$ is within ϵ of β_0.

Although the first step is well-known, for completeness I provide a proof. Because Y is compact and $\mu^n(y|\beta)$ continuous the assumptions A1, A2 and A3 of Andrews (1987) are satisfied so the uniform strong law of large number applies and μ^{nT} almost surely converges uniformly in each component to $E\mu^n|\tilde{F}^n$, hence almost surely in all components.

Suppose μ^{nT} converges uniformly to $E\mu^n|\tilde{F}^n$. For any convergent subsequence of $\hat{\beta}^{nT} \to \hat{\beta}$ and any $\beta \in B$ it is the case that $h(\mu^{nT}(\hat{\beta}^{nT})) \leq h(\mu^{nT}(\beta))$. From uniform convergence $Eh(\mu(\hat{\beta}))|\tilde{F}^n \leq Eh(E\mu(\beta))|\tilde{F}^n$ so $\hat{\beta} \in B^n$. Hence the limit points of $\hat{\beta}^{nT}$ almost surely lie in B^n. □

Appendix III: Exogenous Variables

The space of observations is now a product space: Y, X are compact subsets of a finite-dimensional vector space, and $Z = Y \times X$. All definitions and results in the text remain the same except that everywhere Y in the text should be replaced with Z. Hence $F^n(\beta), \tilde{F}^n, F(\beta)$ are now distributions over Z rather than Y and moments are $\mu(z|\beta)$.

The marginal $\tilde{F}^n_X$ is defined for measurable $A \subseteq X$ in the usual way by $\tilde{F}^n_X(A) \equiv \int_{Y \times A} d\tilde{F}^n(z)$. The new feature is this. The theory rather than specifying a distribution $F^n(\beta)$ over Z instead specifies $G^n(x, \beta)$ a (conditional) distribution over Y. This is assumed to have the measurability property that if $g(y, x)$ is a measurable function then $g_Y(x, \beta) \equiv \int g(y, x) dG^n(x, \beta)$ is measurable. Then $F^n(\beta)$ is then defined through the conditional expectation as the unique measure satisfying

$$\int g(y, x) dF^n(\beta) = \int g_Y(x, \beta) d\tilde{F}^n_X.$$

In particular the probability $F^n(\beta)$ assigns to a measurable set is uniquely defined by the expectation of the indicator function of that set.

The next Proposition shows that uniform convergence of continuous G^n to G together with $\tilde{F}^n_X \to \tilde{F}_X$ implies uniform convergence of F^n to F which is what is required in the definition of approximation. First, a Lemma showing that uniform convergence of G^n implies uniform convergence of g^n_Y

Lemma 7.3. *Suppose* $(x^n, \beta^n) \to (x, \beta)$ *and* $G^n(x^n, \beta^n) \to G(x, \beta)$. *If* $g(y, x)$ *is continuous then* $g^n_Y(x^n, \beta^n) \to g_Y(x, \beta)$, *and* g_Y *is continuous.*

Proof. Suppose that $g(y,x)$ is continuous. Then

$$g_Y^n(x^n,\beta^n) = \int g(y,x^n)dG^n(x^n,\beta^n)$$

$$= \int g(y,x)dG^n(x^n,\beta^n) + \int \left(g(y,x^n) - g(y,x)\right) dG^n(x^n,\beta^n)$$

The first term converges to $g_Y(x,\beta)$ because $G^n(x^n,\beta^n) \to G(x,\beta)$ while the second term converges to zero since the continuity of g implies that $g(\cdot,x^n)$ converges uniformly to $g(\cdot,x)$.

Continuity of $g_Y(x,y)$ follows by applying the first result to the sequence $G^n = G$. □

Proposition 7.4. *Suppose that $G^n(x,\beta)$ is continuous, that $\tilde{F}_X^n \to \tilde{F}_X$, and for all $(x^n,\beta^n) \to (x,\beta)$ it is the case that $G^n(x^n,\beta^n) \to G(x,\beta)$. Then $F^n(\beta^n) \to F(\beta)$ where $\int g(y,x)dF(\beta) = \int g_Y(x,\beta)d\tilde{F}_X(x)$.*

Proof. Suppose $g(y,x)$ is continuous. Then

$$\int g(y,x)dF^n(\beta^n) = \int g_Y^n(x,\beta^n)d\tilde{F}_X^n$$

$$= \int g_Y(x,\beta)d\tilde{F}_X^n + \int \left(g_Y^n(x,\beta^n) - g_Y(x,\beta)\right) d\tilde{F}_X^n.$$

The first term converges to $\int(g(y,x)dF(\beta)$ because $g_Y(x,\beta)$ is continuous and $\tilde{F}_X^n \to \tilde{F}_X$. The second term converges to zero because from Lemma 7.3 $g_Y^n(\cdot,\beta^n)$ converges uniformly to $g_Y(\cdot,\beta)$. □

The generalized and ordinary methods of moments are unchanged and Theorem 5.4 remains true.

References

Andrews, D. W. (1987): "Consistency in nonlinear econometric models: A generic uniform law of large numbers," *Econometrica*: 1465-1471.

Andrews, D. W. (1991): "Heteroskedasticity and autocorrelation consistent covariance matrix estimation," *Econometrica* 817-858.

Billingsley, P. (1968): *Convergence of Probability Measures*, Wiley.

Bó, P. Dal (2005): "Cooperation under the shadow of the future: experimental evidence from infinitely repeated games," *American Economic Review* 95: 1591-1604.

Fehr, Ernst and Klaus M. Schmidt (1999): "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics* 114: 817-868.

Fudenberg, D., and G. Karreskog Rehbinder (2024): "Predicting Cooperation with Learning Models," *American Economic Journal: Microeconomics* 16: 1-32.

Fudenberg, D. and D. K. Levine (1997): "Measuring Players' Losses in Experimental Games," *Quarterly Journal of Economics* 112: 507-536.

Gourieroux, C., A. Monfort and A. Trognon (1984): "Pseudo maximum likelihood methods: Theory," *Econometrica* 681-700.

Hansen, L. P. and T. J. Sargent, T. J. (2001): "Robust control and model uncertainty," *American Economic Review* 91: 60-66.

Hansen, L. P., T. J. Sargent, G. Turmuhambetova and N. Williams (2006): "Robust control and model misspecification," *Journal of Economic Theory*, 128: 45-90.

Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica.*

Leamer, E. E. (1983): "Let's take the con out of econometrics," *American Economic Review* 73: 31-43.

Levine, D. (1983): "A remark on serial correlation in maximum likelihood," *Journal of Econometrics* 23: 337-342.

Levine, D. K. (1986): "Modeling altruism and spitefulness in experiments," *Review of Economic Dynamics* 1: 593-622.

Levine, D. (2024): "Behavioral Mechanism Design as a Benchmark for Experimental Studies," mimeo RHUL.

McKelvey, R. D. and T. R. Palfrey(1995): "Quantal response equilibria for normal form games," *Games and Economic Behavior* 10: 6-38.

Newey, W. K. and D. G. Steigerwald (1997): Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models," *Econometrica*: 587-599.

Nikiforakis, N. and H. T. Normann (2008): "A Comparative Statics Analysis of Punishment in Public-good Experiments," *Experimental Economics* 11: 358-369.

Pearson, K. (1936): "Method of moments and method of maximum likelihood," *Biometrika* 28: 34-59.

Prasnikar, V. and A. E. Roth (1992): "Considerations of fairness and strategy: Experimental data from sequential games," *Quarterly Journal of Economics* 107: 865-888.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara and S. Zamir (1991): "Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study," *American Economic Review* 1068-1095.

Sargent, T. J. (2024): "Macroeconomics After Lucas," tomsargent.com

Watson, J. and C. Holmes(2017): "Approximate models and robust decisions," *Statistical Science* 31.

Stahl II, D. O. and P. W. Wilson (1994): "Experimental evidence on players' models of other players," Journal of Economic Behavior and Organization," 25: 309-327.