A Brief Review of Maximum Likelihood, GMM, and Numerical Tools

Joan Llull

MOVE, Universitat Autònoma de Barcelona
and Barcelona School of Economics

In this introductory notes we quickly review the necessary notions about Maximum Likelihood and GMM estimation methods, which were covered in detail the Econometrics course (first year). Additionally, some numerical methods, which may be useful for some of the problem sets and in other courses, are very briefly introduced. In particular, we provide a brief overview of numerical differentiation, Newton-Raphson optimization, and numerical integration.

I. Maximum Likelihood

A. *The Likelihood Principle*

The maximum likelihood estimator is based on the *likelihood principle*. Following this principle, our estimate of the *true* parameter vector $\boldsymbol{\theta}_0$ is given by the vector $\boldsymbol{\theta}$ that maximizes the likelihood of observing our sample $(\boldsymbol{y}, X) = ((y_1, \boldsymbol{x}_1')', ...(y_N, \boldsymbol{x}_N')')'$. This is opposed to least squares, which instead minimize the sum of squares of residuals. In the case of discrete data, this "likelihood" is given by the probability of drawing the sample, $\Pr[\boldsymbol{y}, X; \boldsymbol{\theta}]$, and in the case of continuous data, it is given by its probability density function (pdf) $f(\boldsymbol{y}, X; \boldsymbol{\theta})$. Without loss of generality, we use the notation $f(\boldsymbol{y}, X; \boldsymbol{\theta})$ for both cases.

The likelihood function, $L_N^*(\boldsymbol{\theta}) \equiv f(\boldsymbol{y}, X; \boldsymbol{\theta})$, is a function that maps a parameter vector $\boldsymbol{\theta}$ and a random sample $(\boldsymbol{y}, X)$ into the probability (or density) that the sample is obtained from the specified model. The term $L_N^*(\boldsymbol{\theta})$ is actually a compact form of $L_N^*(\boldsymbol{\theta}; \boldsymbol{y}, X)$. The likelihood function $L_N^*(\boldsymbol{\theta}) = f(\boldsymbol{y}, X; \boldsymbol{\theta}) = f(\boldsymbol{y}|X; \boldsymbol{\theta}) f(X; \boldsymbol{\theta})$ requires specifying the conditional density of $\boldsymbol{y}$ given X for a given $\boldsymbol{\theta}$, $f(\boldsymbol{y}|X; \boldsymbol{\theta})$, and the marginal density of X, $f(X; \boldsymbol{\theta})$. Under the very general assumption that the distribution of X does not depend on the same set of parameters than the distribution of $\boldsymbol{y}$ —i.e. $f(X; \boldsymbol{\theta}) = f(X)$—, maximizing $L_N^*(\boldsymbol{\theta})$ is equivalent to maximizing the *conditional* likelihood function $L_N(\boldsymbol{\theta}) = f(\boldsymbol{y}|X; \boldsymbol{\theta})$.

Maximizing the likelihood $L_N(\boldsymbol{\theta})$ is equivalent to maximizing the log-likelihood function, $\mathcal{L}_N(\boldsymbol{\theta}) \equiv \ln L_N(\boldsymbol{\theta})$. Let $\{y_i, \boldsymbol{x}_i\}_{i=1}^N$ denote a random sample of independent observations, each with the conditional density function $f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})$. Then $f(\boldsymbol{y}|X; \boldsymbol{\theta}) = \prod_{i=1}^N f(y_i|\boldsymbol{x}_i; \boldsymbol{\theta})$ (given the independence between observations), and

the conditional *log-likelihood* function is:

$$\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{1}$$

B. The Maximum Likelihood Estimator (MLE)

The maximum likelihood estimator (MLE) is defined by the following optimization problem:

$$\hat{\boldsymbol{\theta}}_{ML} \equiv \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}). \tag{2}$$

It is a *fully parametric* estimator because the distribution of the dependent variable y (given independent variables $\boldsymbol{x}$) is completely characterized by a set of parameters and functional forms that we assume to know. This is opposed to *semi-parametric* estimators, which make inference on a set of parameters without need of specifying the entire distribution (e.g. OLS allows to estimate the regression parameters with no need of specifying the distribution of errors), or to *non-parametric* estimators, which make inference about some characteristics of the distribution of the data without parameterizing any element of the distribution (e.g. the sample average of a variable y calculated with the observations that satisfy $x_i = a$ gives a non-parametric estimate of $\mathbb{E}[y|x = a]$). Additionally, the MLE belongs to the very general class of estimators called *extremum estimators*, which are those obtained as a solution of the optimization problem. Specifically, it belongs to the sub-class called *m-estimators*, which maximize an objective function that is an average of subfunctions of the data: $Q_N(\boldsymbol{\theta}) \equiv N^{-1}\sum_{\mathrm{i}=1}^{\mathrm{N}} q(y_i, \boldsymbol{x_i}, \boldsymbol{\theta})$. Another example of an m-estimator is Nonlinear Least Squares (NLS), and an example of an extremum estimator that is not an m-estimator is the Generalized Method of Moments (GMM), which minimizes a quadratic in sample averages.

The solution of the problem in (2) is given by the first order conditions:

$$\frac{1}{N}\frac{\partial\,\mathcal{L}_{\mathrm{N}}(\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}} = \frac{1}{N}\frac{\partial\sum_{\mathrm{i}=1}^{\mathrm{N}}\ln f(y_i|\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}} = \mathbf{0}. \tag{3}$$

C. Asymptotic Properties of the MLE

Identification Assume that there is a *true* parameter vector $\boldsymbol{\theta}_0$ that generates the data. We say that this parameter is identified if there are *no observationally equivalent* parameters. More formally, $\boldsymbol{\theta}_0$ is identified if the *Kullback-Leibler*

inequality is satisfied:

$$\Pr[f(y|\boldsymbol{x};\boldsymbol{\theta}) \neq f(y|\boldsymbol{x};\boldsymbol{\theta}_0)] > 0 \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \tag{4}$$

In words, the true parameter is identified if there is no other parameter vector that generates the same samples with probability one. Put differently, if there exists a parameter vector $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ for which any sample we draw $(\boldsymbol{y}, X)$ has a likelihood equal to $f(\boldsymbol{y}|X;\boldsymbol{\theta}_0)$, then we say that $\boldsymbol{\theta}_0$ is not identified. Identification is an essential assumption for all the analysis below.

Example: A (stupid) example of non-identified parameters in the linear regression context is the following:

$$y = \alpha + \gamma_1 \mathbb{1}\{x \leq 0\} + \gamma_2 \mathbb{1}\{x > 0\} + u, \quad u \sim \mathcal{N}(0, \sigma^2), \tag{5}$$

where $\mathbb{1}\{c\}$ is an indicator that equals 1 if the condition c holds, and 0 otherwise. In this model, α, γ_1 and γ_2 are not identified because we can increase α by any amount a and reduce γ_1 and γ_2 in the same amount and obtain exactly the same likelihood. Put differently:

$$f(\boldsymbol{y}|\boldsymbol{x};\alpha,\gamma_1,\gamma_2,\sigma^2) = f(\boldsymbol{y}|\boldsymbol{x};\alpha+a,\gamma_1-a,\gamma_2-a,\sigma^2). \tag{6}$$

for any sample $(\boldsymbol{y}, \boldsymbol{x})$. As it is impossible to draw a sample $\{y_i, x_i\}_{i=1}^N$ for which (6) is not satisfied, we conclude that α, γ_1 and γ_2 are not separately identified if further restrictions are not imposed (e.g. either α or one γ is normalized to zero).

Regularity conditions Assume: (i) the specified density $f(y|\boldsymbol{x};\boldsymbol{\theta})$ is the data generating process (dgp), and (ii) the support of y does not depend on $\boldsymbol{\theta}$. Then the *regularity conditions*:

$$\mathbb{E}_f\left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \tag{7}$$

and:

$$-\mathbb{E}_f\left[\frac{\partial^2 \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right] = \mathbb{E}_f\left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right], \tag{8}$$

are satisfied. The notation $\mathbb{E}_f$ indicates that the expectations are taken with respect to $f(y|\boldsymbol{x};\boldsymbol{\theta})$; when evaluated at $\boldsymbol{\theta}_0$, these are real expectations, and they are denoted by $\mathbb{E}$. The left hand side of Equation (8) (evaluated at $\boldsymbol{\theta}_0$) is known as the *information matrix*, and the associated regularity condition condition is known as the *information matrix equality*. The regularity conditions are derived in Appendix A, and are useful in the derivation of the main results for MLE.

Consistency Using the first regularity condition and the identification condition (Kullback-Leibler inequality) it can be proven that:

$$\mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta})] < \mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta}_0)] \quad \forall \boldsymbol{\theta} \neq \boldsymbol{\theta}_0. \tag{9}$$

This implies that $\boldsymbol{\theta}_0$ maximizes the population counterpart of the log-likelihood function, $\mathcal{L}_0(\boldsymbol{\theta}) \equiv \mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta})]$. Intuitively, the first regularity condition implies that $\boldsymbol{\theta}_0$ is the solution of the population problem, and the identification condition establishes that this solution is unique (strict inequality).

This result implies consistency. Note that by the Law of Large Numbers (LLN):

$$\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) \underset{p}{\to} \mathbb{E}[\ln f(y|\boldsymbol{x};\boldsymbol{\theta})]. \tag{10}$$

As a result, as $N \to \infty$, (whenever the parameter space Θ is *compact*, and $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta})$ is *measurable* for all $\boldsymbol{\theta}$) the maximizer of $\mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta})$, which is $\hat{\boldsymbol{\theta}}_{ML}$, converges to the maximizer of $\mathcal{L}_0(\boldsymbol{\theta})$, which is $\boldsymbol{\theta}_0$.

Asymptotic distribution To derive the asymptotic distribution of the estimator, we start from the equation that delivers $\hat{\boldsymbol{\theta}}_{ML}$, and expand it with an exact first order *Taylor expansion* around $\boldsymbol{\theta}_0$:

$$0 = \frac{\partial\, \mathcal{L}_{\mathrm{N}}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} = \frac{\partial\, \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2\, \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{11}$$

where $\boldsymbol{\theta}^*$ is a point between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and $\frac{\partial^2\, \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$ is the $K \times K$ Hessian matrix for the log-likelihood evaluated at $\boldsymbol{\theta}^*$. Multiplying the expression by $1/\sqrt{N}$, noting that $1/\sqrt{N} = \sqrt{N} \times (1/N)$, pre-multiplying by the inverse of the Hessian, and making $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ the subject of the equation we obtain:

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left(\frac{1}{N}\frac{\partial^2\, \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)^{-1} \frac{1}{\sqrt{N}}\frac{\partial\, \mathcal{L}_{\mathrm{N}}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\
&= -\left(\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial^2 \ell_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)^{-1} \frac{1}{\sqrt{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}},
\end{aligned} \tag{12}$$

where $\ell_i(\boldsymbol{\theta}) \equiv \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$.

Assume observations are i.i.d., the conditions for consistency and regularity conditions hold, and the population Hessian $\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]$ exists and is non-singular. By the LLN, the first term of the left hand side of Equation (12) converges to:

$$-\left(\frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial^2 \ell_i(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)^{-1} \underset{p}{\to} -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]^{-1}, \tag{13}$$

as consistency implies that $\boldsymbol{\theta}^* \underset{p}{\to} \boldsymbol{\theta}_0$ by construction (given $\boldsymbol{\theta}^* \in [\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0]$ and $\hat{\boldsymbol{\theta}} \underset{p}{\to} \boldsymbol{\theta}_0$). By the CLT, the second term of Equation (12) satisfies:

$$\frac{1}{\sqrt{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\partial \ell_i(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \underset{d}{\to} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right]\right). \tag{14}$$

Finally, using the Cramer Theorem, we establish the result:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \underset{d}{\to} \mathcal{N}\left(\mathbf{0}, \Omega_0\right), \tag{15}$$

where, given the information matrix equality, Ω_0 satisfies:

$$\begin{aligned} \Omega_0 &= \mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right] \mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} \\ &= -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1} = \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right]^{-1}. \end{aligned} \tag{16}$$

These two versions of the variance-covariance matrix are very useful. The Hessian version of the formula indicates that the precision of the estimation depends on the curvature of the likelihood function around $\boldsymbol{\theta}_0$. Additionally, since it corresponds to the inverse of the (Fisher) information matrix, it is the *Cramer-Rao* lower bound, which is the lowest variance of unbiased estimators in finite samples. The Jacobian product version is computationally handy because the partial derivatives are often computed in the estimation process, and the computation of the Hessian, which may be time consuming, is avoided.

II. Generalized Method of Moments (GMM)

A. *General Formulation*

Let $\boldsymbol{\theta}$ be the parameter vector of interest, defined by the set of moments (or orthogonality conditions):

$$\mathbb{E}[\psi(\boldsymbol{w}; \boldsymbol{\theta})] = \mathbf{0}, \tag{17}$$

where $\boldsymbol{w}$ is a (vector) random variable, and $\psi(\cdot)$ is a vector function such that $\dim(\psi) \geq \dim(\boldsymbol{\theta})$. Therefore, Equation (17) specifies $\dim(\psi)$ moment conditions. ***Example:*** Consider a regression model. The parameter vector, $\boldsymbol{\beta}$, is such that:

$$\mathbb{E}[\boldsymbol{z}u] = \mathbf{0}, \tag{18}$$

where:

$$u \equiv y - f(\boldsymbol{x}; \boldsymbol{\beta}), \quad \text{and} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}), \tag{19}$$

and $\dim(\boldsymbol{z}) \geq \dim(\boldsymbol{\beta})$.

B. *Estimation*

Consider a random sample with N observations, $\{\boldsymbol{w}_i\}_{i=1}^N$. The GMM estimation is based on the sample analog of Equation (17):

$$\boldsymbol{b}_N(\boldsymbol{\theta}) \equiv \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \psi(\boldsymbol{w}_i;\boldsymbol{\theta}). \tag{20}$$

The GMM estimator is given by the value of $\boldsymbol{\theta}$ that minimizes the quadratic distance of $b_N(\boldsymbol{\theta})$ from zero:

$$\hat{\boldsymbol{\theta}}_{GMM} \equiv \arg\min_{\theta\in\Theta} \boldsymbol{b}_N(\boldsymbol{\theta})' W_N \boldsymbol{b}_N(\boldsymbol{\theta}), \tag{21}$$

where W_N is a squared semi-positive definite weighting matrix that satisfies the rank condition $\text{rank}(W_N) \geq \dim(\boldsymbol{\theta})$. Note that, if the problem is just-identified, this is when $\dim(\psi) = \dim(\boldsymbol{\theta})$, the weighting matrix becomes irrelevant, and the GMM estimator satisfies:

$$\boldsymbol{b}_N(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}. \tag{22}$$

Example: Building on the previous example, consider the linear regression model, i.e. $f(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$. Then:

$$\boldsymbol{b}_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{\mathrm{i=1}}^{\mathrm{N}} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{1}{N} Z'(\boldsymbol{y} - X\beta), \tag{23}$$

and $\hat{\boldsymbol{\beta}}_{GMM}$ satisfies:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GMM} &= \arg\min_{\beta} N^{-2}(\boldsymbol{y} - X\beta)' Z W_N Z'(\boldsymbol{y} - X\beta) \\ &= (X'ZW_NZ'X)^{-1} X'ZW_NZ'\boldsymbol{y}, \end{aligned} \tag{24}$$

which equals the 2SLS estimator when $W_N = (Z'Z)^{-1}$.

C. *Asymptotic Properties*

As GMM is an extremum estimator, the general asymptotic results for this type of estimators hold. Thus, conditions and derivations are similar to those for MLE.

Consistency Assume the parameter space $\Theta \in \mathbb{R}^K$ is compact, the criterion function converges in probability to its population counterpart, i.e. $W_N \boldsymbol{b}_N(\boldsymbol{\theta}) \underset{p}{\to} W_0\, \mathbb{E}[\psi(\boldsymbol{w};\boldsymbol{\theta})]$, and the parameter vector is identified, i.e. $\boldsymbol{\theta}_0$ is the only solution of the population problem $W_0\, \mathbb{E}[\psi(\boldsymbol{w};\boldsymbol{\theta})] = 0$. Then, $\hat{\boldsymbol{\theta}}_{GMM} \underset{p}{\to} \boldsymbol{\theta}_0$.

Asymptotic distribution Assume consistency conditions are satisfied; $\boldsymbol{\theta}$ is in the interior of Θ; $\psi(\boldsymbol{w};\boldsymbol{\theta})$ is once differentiable with respect to $\boldsymbol{\theta}$; $D_N(\boldsymbol{\theta}) \equiv \partial \boldsymbol{b}_N(\boldsymbol{\theta})/\boldsymbol{\theta}'$ converges in probability to $D_0(\boldsymbol{\theta})$, where $D_0(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; for $D_0 \equiv D_0(\boldsymbol{\theta}_0)$, the matrix $D_0' W_0 D_0$ is non-singular; and $\sqrt{N}\boldsymbol{b}_N(\boldsymbol{\theta}_0) \underset{d}{\to} \mathcal{N}(0, V_0)$, with $V_0 \equiv \mathbb{E}[\psi(\boldsymbol{w};\boldsymbol{\theta}_0)\psi(\boldsymbol{w};\boldsymbol{\theta}_0)']$. Following similar steps as in Section I.C, we can show that $\sqrt{N}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \underset{d}{\to} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$, where:

$$\boldsymbol{\Omega}_0 = (D_0' W_0 D_0)^{-1} D_0' W_0 V_0 W_0 D_0 (D_0' W_0 D_0)^{-1}. \tag{25}$$

Optimal weighting matrix Even though any semi-positive definite weighting matrix that satisfies the rank condition provides a consistent estimate of $\boldsymbol{\theta}_0$, not all of them form an efficient estimator. Efficiency is achieved with any W_N that implies $W_0 = \kappa V_0^{-1}$, for a positive κ. This includes $W_N = V_0^{-1}$ (unfeasible), but also $W_N = \hat{V}_N^{-1}$, where $\hat{V}_N$ is any consistent estimator of V_0. In practice, the Optimal GMM estimator is implemented in two steps:

1) Obtain $\hat{\boldsymbol{\theta}}_{GMM}(W_N^0)$ for an initial guess W_N^0.

2) Re-estimate using $\widehat{W}_{opt} \equiv (\sum_{\mathrm{i}=1}^{\mathrm{N}} \psi(\boldsymbol{w}_i; \hat{\boldsymbol{\theta}}_{GMM}(W_N^0))\psi(\boldsymbol{w}_i; \hat{\boldsymbol{\theta}}_{GMM}(W_N^0))')^{-1}$ as the new weighting matrix.

III. Numerical Methods

This section briefly reviews some of the numerical methods that you will use in the problem sets. In all cases, there are many alternative methods that could be implemented. We do not review all of them. See Judd (1998) for detailed descriptions of many of these algorithms.

A. Differentiation

Although analytical differentiation is always preferred when possible, we might be interested in numerical differentiation either because of the absence of closed form solutions, or to avoid tedious and complicated derivations. Numerical differentiation is based on the definition of a derivative:

$$f'(x) = \lim_{\epsilon \to 0} \frac{f(x+\epsilon) - f(x)}{\epsilon}. \tag{26}$$

This suggests the formula:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \tag{27}$$

for a small h (e.g. 10^{-6}). The expression in Equation (27) is known as a *one-sided* differential. A more accurate alternative is the so-called *two-sided* differential:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}, \tag{28}$$

even though the extra accuracy comes at the expense of additional evaluations, which increase the computational burden. Whenever the argument of the function is a vector, a gradient $\nabla_f(\boldsymbol{x})$ needs to be computed. Each element of $\nabla_f(\boldsymbol{x})$ is calculated by perturbing one of the elements of $\boldsymbol{x}$ through a the small scalar h, leaving all other elements to the baseline value. The additional burden implied by two-sided differentiation becomes more salient.

B. Newton-Raphson Optimization

The original version of the Newton-Raphson method (a.k.a. as Newton's method) was conceived to find the roots of a given function. The extension to optimization is natural, as optimization consists of finding roots to the first order conditions. The method is an iterative algorithm that approximates the function in a given point x_n by its tangent line, and then finds the intercept of this line with respect to the horizontal axis to update the guess and iterate again. The new point x_{n+1} will be an improvement as long as the function is globally convex or globally concave.

More formally, assume that we want to find the root of $f(x)$ (which is at least once differentiable). We know that the derivative of the function must be equal to the slope of the hypotenuse of the triangle $((x_n, 0), (x_n, f(x_n)), (x_{n+1}, 0))$. Hence:

$$\frac{f(x_n) - 0}{x_n - x_{n+1}} = f'(x_n) \Rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{29}$$

This function describes an iterative procedure that we can execute starting from an initial guess until reaching convergence, i.e. until $|x_{n+1} - x_n| < \epsilon$ for a small ϵ.

Given that, as noted above, optimization consists of finding roots for the first order condition, the Newton step for minimization takes the form of:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}. \tag{30}$$

If the function is $f : \mathbb{R}^K \to \mathbb{R}$, at least twice differentiable, the Newton step is:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - [H_f(\boldsymbol{x}_n)]^{-1} \nabla_f(\boldsymbol{x}_n), \tag{31}$$

where H_f indicates the Hessian of f, and ∇_f is the gradient. When the Hessian is computationally too demanding, there are alternative methods (called Quasi-Newton methods) that avoid its computation.

C. Integration

Numerical integration (a.k.a. quadrature) is an approximation to the value of an integral between two points. In general, the integrand is evaluated at a finite set of points, called integration points, and the integral is approximated by a weighted sum of these values. The integration and weights depend on the specific method used, and on the accuracy required from the approximation. Simple (and less precise rules) include the midpoint rule, the trapezoidal rule, and the Simpson's rule. If the integrand is smooth, Gaussian quadrature formulas are typically more accurate. An alternative to deterministic quadrature methods is Monte Carlo integration, that uses uniformly generated random numbers as integration points. In particular, the method is as simple as drawing a set of random points at which the function is evaluated, and then, averaging function evaluations.

References

Arellano, Manuel (2003), Panel Data Econometrics, Oxford University Press, Appendix A.

Cameron, A. Colin and Pravin K. Triverdi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press, Chapters 5 and 6.

Judd, Kenneth L. (1998), *Numerical Methods in Economics*, The MIT Press, Chapters 4, 7 and 8.

Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Chapter 13.

Appendix A: Derivation of the Regularity Conditions

The derivation of the two conditions is quite simple. First, note that the density $f(y|\boldsymbol{x};\boldsymbol{\theta})$ integrates to one, which implies:

$$\int f(y|\boldsymbol{x};\boldsymbol{\theta})dy = 1 \quad \Rightarrow \quad \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y|\boldsymbol{x};\boldsymbol{\theta})dy = \mathbf{0}. \tag{A1}$$

Given assumption (ii), the support of y does not depend on $\boldsymbol{\theta}$, so we can swap integration and differentiation, which yields:

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y|\boldsymbol{x};\boldsymbol{\theta})dy \underset{(ii)}{=} \int \frac{\partial f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy. \tag{A2}$$

Taking into account that the relationship between the partial derivatives of the likelihood and the log-likelihood is given by:

$$\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{f(y|\boldsymbol{x};\boldsymbol{\theta})}\frac{\partial f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \Rightarrow \quad \frac{\partial f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{x};\boldsymbol{\theta}), \tag{A3}$$

we substitute this into Equation (A2) to obtain the first regularity condition:

$$\int \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{x};\boldsymbol{\theta}) dy = \mathbb{E}_f \left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] = \mathbf{0}. \tag{A4}$$

Equation (8) is derived from the derivative of the first regularity condition (Equation (7)) with respect to $\boldsymbol{\theta}'$:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\theta}'} \int \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{x};\boldsymbol{\theta}) dy \underset{(ii)}{=} \int \frac{\partial}{\partial \boldsymbol{\theta}'} \left(\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(y|\boldsymbol{x};\boldsymbol{\theta}) \right) dy = \\
&= \int \left(\frac{\partial^2 \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(y|\boldsymbol{x};\boldsymbol{\theta}) + \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right) dy = \\
&\underset{(A3)}{=} \int \left(\frac{\partial^2 \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(y|\boldsymbol{x};\boldsymbol{\theta}) + \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(y|\boldsymbol{x};\boldsymbol{\theta}) \right) dy = \\
&= \mathbb{E}_f \left[\frac{\partial^2 \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] + \mathbb{E}_f \left[\frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \ln f(y|\boldsymbol{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] = \mathbf{0}.
\end{aligned} \tag{A5}$$

Equation (8) is then trivially derived from Equation (A5). ■