Chapter 11

EXTREMUM ESTIMATORS

One of the common features across many estimators that are widely used in application such as ordinary least squares, instrumental variables, GMM, and maximum likelihood estimators, is that they are obtained by minimizing or maximizing an objective function. These estimators are called extremum estimators, or optimization estimators. This chapter explains a unified framework for this class of estimators.

11.1 Asymptotic Properties of Extremum Estimators

Let $\{\mathbf{x}_t : t = 1, 2, \cdots T\}$ be a vector stochastic process, $\mathbf{b}_0$ be a p-dimensional vector of parameters to be estimated, and $J(\mathbf{b})$ be a real-valued objective function. For notational simplicity, the dependency of $J(\mathbf{b})$ on $\{\mathbf{x}_t : t = 1, 2, \cdots, T\}$ is suppressed. An *extremum estimator* is a vector of parameters, $\mathbf{b_T}$, which minimizes the objective function, $J_T(\mathbf{b})$, with respect to $\mathbf{b}$. Under general regularity conditions, an extremum estimator is consistent and asymptotically normally distributed.[1]

There are two important assumptions that ensure the consistency and asymptotic normality of extremum estimators: convergence and identification.

[1]See the Appendix of Chapter 9 for a proof of consistency.

11.1.1 Convergence

The convergence assumption is that $J_T(\mathbf{b})$ converges with probability one to some deterministic function $J_0(\mathbf{b})$ as $T \to \infty$ for all admissible values of $\mathbf{b}$. Convergence may take different forms such as uniform convergence and convergence in probability.

11.1.2 Identification

The identification assumption is that $\mathbf{b}_0$ is the unique minimizer of $J_0(\mathbf{b})$.

11.2 Two Classes of Extremum Estimators

There are two classes of extremum estimators, *classical minimum distance estimators* and *M-estimators*.

11.2.1 Minimum Distance Estimators

An extremum estimator is a minimum distance estimator if the objective function is a quadratic function:

$$J_T(\mathbf{b}) = f_T(\mathbf{b})'\mathbf{W}_T f_T(\mathbf{b}), \tag{11.1}$$

where $f(\cdot)$ is a q-dimensional vector of functions and $\mathbf{W}_T$ is a sequence of matrix that satisfies

$$\lim_{T\to\infty} \mathbf{W}_T = \mathbf{W}_0 \tag{11.2}$$

with probability one for a positive definite matrix $\mathbf{W}_0$. The matrices $\mathbf{W}_T$ and $\mathbf{W}_0$ are called the distance, or weighting, matrix.

A prominent example of the minimum distance estimator is the GMM estimator. In the GMM, the sample mean is used for $f_T(\mathbf{b})$, and the law of large number for the sample mean ensures convergence.

11.2.2 M-Estimators

An extremum estimator is an *M-estimator* if the objective function is a sample average:

$$Q_T(\mathbf{b}) = \frac{\mathbf{1}}{\mathbf{T}} \sum_{\mathbf{t=1}}^{\mathbf{T}} \mathbf{m}(\mathbf{x_t}), \tag{11.3}$$

where $m(\cdot)$ is a real-valued function. The maximum likelihood (ML) estimator is a leading example of the M-estimator. Suppose $\{\mathbf{x}_t\}$ is an i.i.d. process with a known density function $f(\mathbf{x}_t; \mathbf{b_0})$ where $\mathbf{b_0}$ is an unknown true parameter vector. The joint density of $\{\mathbf{x}_t\}$ is given by

$$f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T; \mathbf{b_0}) = \prod_{t=1}^{T} f(\mathbf{x}_t; \mathbf{b}_0). \tag{11.4}$$

If we replace $\mathbf{b}_0$ with some arbitrary (random?) value $\mathbf{b}$, and interpret the density as a function of $\mathbf{b}$, it is called the *likelihood function*. The ML estimator for $\mathbf{b}_0$ is a parameter vector $\mathbf{b}$ that maximizes the likelihood function. Since the log transformation is a monotone transformation, maximizing the likelihood function is equivalent to minimizing the following:

$$-\log f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T; \mathbf{b_0}) = -\sum_{t=1}^{T} f(\mathbf{x}_t; \mathbf{b}_0). \tag{11.5}$$

11.3 Examples of Minimum Distance Estimators

11.3.1 Two-Step Minimum Distance Estimators

Another example of the minimum distance estimator is a *two-step minimum distance estimator*. Suppose $\mathbf{c_0}$ is the true values of some parameters of interest. In the first step, a consistent estimator for $\mathbf{c_0}$, $\mathbf{c}_T$, is obtained. In the second step, the minimum

distance method is used to estimate another set of parameters based on $\mathbf{c}_T$ from the first step.

One application of the two-step minimum distance estimation has an unrestricted estimator as $\mathbf{c}_T$ and uses the minimum distance estimation to impose restrictions on $\mathbf{c}_0$. Suppose $\mathbf{c}_T$ is an unrestricted estimator for a $(p+s)$-dimensional vector of parameters $\mathbf{c}_0$. Consider nonlinear restrictions

$$\phi(\mathbf{b}_0) = \mathbf{c}_0, \tag{11.6}$$

where $\mathbf{b}_0$ is a p-dimensional vector of parameters. The minimum distance estimator, $\mathbf{b}_T$, minimizes

$$J_T(\mathbf{b}) = \{\phi(\mathbf{b}) - \mathbf{c}_T\}'\mathbf{W}_T\{\phi(\mathbf{b}) - \mathbf{c}_T\}, \tag{11.7}$$

where $\mathbf{W}_T$ is a positive definite distance matrix and converges to some positive definite matrix $\mathbf{W}_0$ with probability one. As in the GMM, the optimal distance matrix is $\mathbf{W} = \mathbf{\Omega}^{-1}$ and $TJ_T(\mathbf{b}_T)$ has an (asymptotic) chi-square distribution with s degrees of freedom. The null hypothesis (11.6) is rejected when this statistic exceeds the critical value from a chi-square distribution. See Altug and Miller (1990) and Atkeson and Ogaki (1996) for empirical applications.

11.3.2 Two-Step Minimum Distance Estimation with Impulse Responses

Another application of the two-step minimum distance estimator is the estimation of parameters in a theoretical model by matching the model's theoretical impulse response functions with empirical impulse response functions estimated by vector autoregressions (VAR). Denoting a vector of model parameters by $\boldsymbol{\beta}$, the optimal

estimators are chosen so as to minimize the quadratic distance between empirical impulse responses, denoted by $\hat{\Psi}$, and the model-implied impulse responses:

$$\min_{\boldsymbol{\beta}} \left[\hat{\Psi} - \Psi(\boldsymbol{\beta})\right]' \boldsymbol{\Sigma}^{-1} \left[\hat{\Psi} - \Psi(\boldsymbol{\beta})\right], \tag{11.8}$$

where $\Psi(\boldsymbol{\beta})$ denotes the mapping from $\boldsymbol{\beta}$ to the model impulse response functions, and $\boldsymbol{\Sigma}$ is a diagonal matrix whose diagonal elements are sample variances of the $\hat{\Psi}$'s.

Sbordone (2002) and Sbordone (2005) apply this method to estimate the degree of price stickiness from the NKPC. The so-called Calvo (1983) parameter measures the probability that a firm does not change its price in a given period. Letting θ denote this probability, the average number of periods for which a price remains unchanged is $(1-\theta)\sum_{k=0}^{\infty} k\theta^{k-1} = 1/(1-\theta)$. Magnusson and Mavroeidis (2009) develop the identification robust minimum distance estimator with similar ideas as the identification robust GMM estimator. However, their confidence sets indicate that the minimum distance estimation applied to the NKPC is subject to the weak identification problem. For example, their 95% confidence interval for the average price duration has a lower bound of around 3.3 quarters and an upper bound of infinity.

A classic method to estimate θ is the single-equation GMM using the NKPC (see, for example, Galí and Gertler (1999) and Eichenbaum and Fisher (2007)). In Galí and Gertler (1999), θ is estimated to be around 0.8, implying the average price duration of 5 quarters. However, as surveyed by Kleibergen and Mavroeidis (2009), this estimation method is also subject to the weak identification problem. The 95% confidence interval for the average price duration using their recommended method has a lower bound of two quarters and an upper bound of infinity. Since the lower bound obtained from the minimum distance method is sharper than that from the

GMM, the minimum distance method outperforms the GMM when applied to a single equation using the NKPC.

Christiano, Eichenbaum, and Evans (2005) apply the two-step minimum distance method to a system of equations from their DSGE model to investigate the role of nominal rigidities in generating the observed persistent responses of inflation and output to a monetary policy shock. They first estimate the VAR impulse responses of 8 key macroeconomic variables using the post-war U.S. data. Let $\mathbf{Y}_{1t}$ be a vector of observations on real GDP, real consumption, GDP deflator, real investment, and real wage, R_t denote the federal funds rate, and $\mathbf{Y}_{2t}$ be a vector of real profits and the growth rate of M2. These variables are stacked as $\mathbf{Y}_t = [\ \mathbf{Y}_{1t}' \quad R_t \quad \mathbf{Y}_{2t}'\ ]'$. This ordering ensures that the monetary policy shock is identified by two identifying assumptions. First, the variables in $\mathbf{Y}_{1t}$ are assumed not to respond contemporaneously to the monetary policy shock, and second, the federal funds rate does not depend on the current values of the variables in $\mathbf{Y}_{2t}$. Using the first 25 estimated coefficients of each impulse response as elements of $\hat{\mathbf{\Psi}}$ in (16.11), model parameters are estimated as a solution to (16.11). Their estimate of θ is 0.6 in the benchmark model, implying the average price duration of 2.5 quarters. Because they apply the method to a system of equations rather than a single equation, their system may be well identified. This is an important topic for further research.[2]

[2]Kim and Ogaki (2009) estimate the Calvo parameter in an exchange rate model with the Taylor rule without the NKPC. In their estimation for θ, there is a substantial efficiency gain by applying the GMM to a system of equations rather than to a single equation. We expect an analogous substantial efficiency gain for the minimum distance estimation.

11.3.3 Minimum Distance to Estimate Data Statistics

Another application of the minimum distance method in the DSGE literature is to estimate various statistics of model variables such as mean, standard deviation, correlation, and autocorrelation. Although the GMM may be used, minimum distance may be more convenient.

Consider two stationary variables, x_t and y_t. Suppose we want to their estimate population moments, $\mathbf{b}_0 = (E(x_t), E(x_t^2), E(y_t), E(y_t^2), E(x_t y_t), E(x_t x_{t-1}))$. Let $\mathbf{x}_t = (x_t, y_t)$ and $f(\mathbf{x}_t, \mathbf{b}) = (x_t, x_t^2, y_t, y_t^2, x_t y_t, x_t x_{t-1})' - \mathbf{b}$, where $f(\mathbf{x}_t, \mathbf{b})$ is a disturbance defined at time t. The GMM minimizes a quadratic form of the sample average of $f(\mathbf{x}_t, \mathbf{b})$, to obtain an estimate of $\mathbf{b}_0$, $\mathbf{b}_T$, and an estimate of covariance matrix of $T^{\frac{1}{2}}(\mathbf{b}_T - \mathbf{b}_0)$.

To obtain the standard errors of estimated statistics that are nonlinear functions of $\mathbf{b}_0$ such as standard deviations, correlations, and autocorrelations, one can use the delta method explained in Proposition 5.8. For example, let $a(\mathbf{b}_0)$ denote the standard deviation of x_t, $a(\mathbf{b}_0) = \sqrt{\text{var}(x_t)} = (E(x_t^2) - E(x_t)^2)^{\frac{1}{2}}$, and $a(\mathbf{b}_T)$ be a consistent estimator of $a(\mathbf{b}_0)$. By the delta method, $\sqrt{T}(a(\mathbf{b}_T) - a(\mathbf{b}_0))$ has an approximate normal distribution with variance $\mathbf{d}(\mathbf{b}_0) Cov(\mathbf{\Omega}^{-1}) \mathbf{d}(\mathbf{b}_0)'$ in a large sample where $\mathbf{d}(\mathbf{b}_0)$ is the derivative of $a(\cdot)$ evaluated at $\mathbf{b}_0$.

In the GMM, while parameters may enter moment conditions nonlinearly, sample moments may not because the moment conditions may not be equal to zero in that case. For example, in order to estimate the variance of x_t in the above example, the moment condition would be $b - (x_t - \bar{x})^2$ where b is the variance to be estimated and $\bar{x}$ is the sample mean. However, because the sample mean enters the moment condition in a nonlinear way, $E(b - (x_t - \bar{x})^2)$ is not equal to zero, which prevents the

GMM estimation.

By contrast, in the minimum distance estimation, sample moments may enter moment conditions in nonlinear ways. For example, Ambler, Cardia, and Zimmermann (2004) (section 3) estimate a pair of correlations

$$\bar{\rho_{1t}} = \frac{(x_1 - \bar{x_1})(x_2 - \bar{x_2})}{\bar{\sigma}_1 \bar{\sigma}_2} \tag{11.9}$$

and

$$\bar{\rho_{2t}} = \frac{(x_3 - \bar{x_3})(x_4 - \bar{x_4})}{\bar{\sigma}_3 \bar{\sigma}_4}, \tag{11.10}$$

where $\bar{x_i}$ and $\bar{\sigma}_i$ are the sample mean and variance of x_i. The optimal estimators are obtained by minimizing

$$\{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t)\}' \mathbf{W}_T \{\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\rho} - \bar{\boldsymbol{\rho}}_t)\} \tag{11.11}$$

where $\boldsymbol{\rho}$ is a (2×1) vector of population correlations of x_{it} for $i = 1, 2, 3, 4$ and $\bar{\boldsymbol{\rho}}_t = [\ \bar{\rho_{1t}} \quad \bar{\rho_{2t}}\ ]'$.

Although this setup resembles the GMM, it cannot be embedded in the standard GMM framework because the sample mean and variance enter the moment conditions nonlinearly. Instead, this is a minimum distance estimator.

The minimum distance estimator can be used to estimate a DSGE model by matching the model-implied moments with empirical moments in a similar way as GMM while allowing the sample mean to enter moment conditions nonlinearly. An application can be found in García-Cicco, Pancrazi, and Uribe (2009).

11.4 The Kalman Filter

We introduced the ML estimator for an i.i.d. process. However, this i.i.d. assumption rarely holds in time series data. In linear models with time dependence, the likelihood

function can be evaluated using a recursive linear algorithm called the Kalman filter (Kalman, 1960). The Kalman filter estimates an evolution of unobserved variable(s) of interest in a discrete-time dynamic system by sequentially updating a linear projection using current observations. Because this filtering process minimizes the mean squared prediction error, it yields an optimal estimator among the class of linear projections. Due to its accuracy and practicality, various extensions of the Kalman filter have been developed and applied in a broad area of study. In econometric, it is used to construct exact finite-sample forecasting, evaluate the exact likelihood function, and estimate parameters in ARMA models or time-varying parameters in linear regressions, just to name a few examples.

In order to formulate the Kalman filter algorithm, the process of interest is modeled in a set of linear equations called the *state-space representation*. This equation system characterizes the relationship between observed and unobserved variables. Let $\mathbf{x}_t$ be an r-dimensional vector of unobserved variables, $\mathbf{y}_t$ be an n-dimensional vector of observed variables, and $\mathbf{z}_t$ be a k-dimensional vector of exogenous variables. Suppose $\mathbf{y}_t$ depends linearly on $\mathbf{x}_t$ and $\mathbf{z}_t$:

$$\mathbf{y}_t = \mathbf{A}' \cdot \mathbf{z}_t + \mathbf{H}' \cdot \mathbf{x}_t + \mathbf{e}_t, \tag{11.12}$$

where $\mathbf{e}_t$ is $(n \times 1)$ vector white noise with $E(\mathbf{e}_t \mathbf{e}_j') = \mathbf{R}$ for $t = j$ and $\mathbf{0}$ otherwise, and $\mathbf{A}'$ and $\mathbf{H}'$ are $(n \times k)$ and $(n \times r)$ matrices of parameters, respectively.

The unobserved vector $\mathbf{x}_t$, called the *state vector*, is assumed to evolve according to a linear stochastic difference equation

$$\mathbf{x}_{t+1} = \mathbf{F} \cdot \mathbf{x}_t + \mathbf{u}_{t+1}, \tag{11.13}$$

where $\mathbf{u}_{t+1}$ is also $(r \times 1)$ vector white noise with $E(\mathbf{u}_t \mathbf{u}_j') = \mathbf{Q}$ for $t = j$ and $\mathbf{0}$

otherwise, and $\mathbf{F}$ is an $(r \times r)$ matrix of parameters. The disturbances $\mathbf{e}_t$ and $\mathbf{u}_t$ are assumed to be independent of each other at all lags, $E(\mathbf{e}_t\mathbf{u}_j') = 0$ for all t and j, and the initial state $\mathbf{z}_1$ is uncorrelated with any realizations of $\mathbf{e}_t$ and $\mathbf{u}_t$, $E(\mathbf{e}_t\mathbf{z}_1') = 0$ and $E(\mathbf{u}_t\mathbf{z}_1') = 0$ for $t = 1, \cdots, T$. Together with the state equation (11.13), the latter assumption implies that $\mathbf{e}_t$ and $\mathbf{u}_t$ are uncorrelated with all lagged values of $\mathbf{x}_t$: $E(\mathbf{e}_t\mathbf{x}_j') = 0$ and $E(\mathbf{u}_t\mathbf{x}_j') = 0$ for $j = t-1, t-2, \cdots, 1$.

Equation (11.12) is called the *observation equation*, and equation (11.13) the *state equation*. Together, they comprise the state-space representation of the dynamics of $\mathbf{y}$.

The Kalman filter recursively generates least square forecasts of the unobserved state vector $\mathbf{x}_t$ as a linear function of the observed data $\mathbf{y}_t$ and $\mathbf{z}_t$. Let $\hat{\mathbf{x}}_{t+1|t} \equiv \hat{E}(\mathbf{x}_{t+1}|\mathbf{\Omega}_t)$ denote the best forecasts of $\mathbf{x}_{t+1}$ based on the data available at time t, $\mathbf{\Omega}_t \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \cdots, \mathbf{y}_1', \mathbf{z}_t', \mathbf{z}_{t-1}', \cdots, \mathbf{z}_1')$. The accuracy of each forecast is measured by an associated $(r \times r)$ error covariance matrix, $\mathbf{P}_{t+1|t} \equiv E[(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1|t})(\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1|t})']$.

In order to initiate the recursive process, the unconditional mean of the initial state $\hat{\mathbf{x}}_{1|0}$ and its covariance $\mathbf{P}_{1|0}$ must be chosen. If the eigenvalues of $\mathbf{F}$ are inside the unit circle, $\hat{\mathbf{x}}_{1|0}$ is simply set equal to $\mathbf{0}$ with an associated covariance matrix whose column vectors are given by $\text{vec}(\mathbf{P}_{1|0}) = [\mathbf{I}_{r^2} - (\mathbf{F} \times \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q})$. Otherwise, the researcher's best guess of $\mathbf{x}_{1|0}$ can be used as $\hat{\mathbf{x}}_{1|0}$, and a positive definite matrix that summarizes the confidence in this guess is used as $\mathbf{P}_{1|0}$.

Suppose we have data on $(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T, \mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)$. For simple illustration, assume that the matrices $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ are known and constant. Given $\hat{\mathbf{x}}_{1|0}$ and $\mathbf{P}_{1|0}$, the linear projection of $\hat{\mathbf{x}}_{t+1|t}$ and associated covariance of this forecast $\mathbf{P}_{t+1|t}$

are iterated on

$$\begin{array}{l}\hat{\mathbf{x}}_{t+1|t} = \mathbf{F}\hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{A}'\mathbf{z}_t - \mathbf{H}'\hat{\mathbf{x}}_{t|t-1}), \\ \mathbf{P}_{t+1|t} = \mathbf{F}[\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}]\mathbf{F}' + \mathbf{Q},\end{array} \tag{11.14}$$

for $t = 1, 2, \cdots, T$, where $\mathbf{K}_t \equiv \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}$ is called the *Kalman gain*. That $\mathbf{K}_t$ depends negatively on $\mathbf{R}$ implies that, when computing the projection for next period, the Kalman filter attaches a smaller (larger) weight to the observation the larger (smaller) the noise in the observed data is (and hence the larger (smaller) $\mathbf{R}$ is).

The previous period's projections are updated based on the current realization of the observable as follows:

$$\begin{array}{l}\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{F}^{-1}\mathbf{K}(\mathbf{y}_t - \mathbf{A}'\mathbf{z}_t - \mathbf{H}'\hat{\mathbf{x}}_{t|t-1}), \\ \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}.\end{array} \tag{11.15}$$

Notice that equations (16.61) and (16.62) are related by:

$$\begin{aligned}\hat{\mathbf{x}}_{t+1|t} &= \mathbf{F}\hat{\mathbf{x}}_{t|t}, \\ \mathbf{P}_{t+1|t} &= \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q}.\end{aligned}$$

Thus, the Kalman filter repeats a project-and-update cycle in which it makes projections $\hat{\mathbf{x}}_{t|t-1}$, updates these projections based on the current observations to get $\hat{\mathbf{x}}_{t|t}$, and uses them to obtain next projections $\hat{\mathbf{x}}_{t+1|t}$. This recursive nature implies that all the necessary information is contained in previous forecasts and information sets, and hence the filtering does not require all the previous data to be stored and re-processed in each estimation step. This is one of the appealing features of the Kalman filter for practical implementations.

Finally, the forecast of $\mathbf{y}_{t+1}$ is obtained as follows. The exogeneity assumption of $\mathbf{z}_t$ implies that it contains no information about $\mathbf{x}_t$ beyond what is contained in

the $t-1$ information set $\mathbf{\Omega}_{t-1} \equiv (\mathbf{y}_{t-1}^{'}, \mathbf{y}_{t-2}^{'}, \cdots, \mathbf{y}_{1}^{'}, \mathbf{z}_{t-1}^{'}, \mathbf{z}_{t-2}^{'}, \cdots, \mathbf{z}_{1}^{'})$. Hence,

$$\hat{E}(\mathbf{x}_t|\mathbf{z}_t, \mathbf{\Omega}_{t-1}) = \hat{E}(\mathbf{x}_t|\mathbf{\Omega}_{t-1}) = \hat{\mathbf{x}}_{t|t-1}.$$

From the observation equation (11.12) and by the law of iterated projections, the forecast of $\mathbf{y}_{t+1}$ is given by

$$\begin{aligned} \hat{\mathbf{y}}_{t+1|t} &\equiv \hat{E}(\mathbf{y}_{t+1}|\mathbf{z}_{t+1}, \mathbf{\Omega}_t) \\ &= \mathbf{A}'\mathbf{z}_{t+1} + \mathbf{H}'\hat{E}(\mathbf{x}_{t+1}|\mathbf{z}_{t+1}, \mathbf{\Omega}_t) \\ &= \mathbf{A}'\mathbf{z}_{t+1} + \mathbf{H}'\hat{\mathbf{x}}_{t+1|t}, \end{aligned}$$

with error covariance

$$E[(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})'] = \mathbf{H}'\mathbf{P}_{t+1|t}\mathbf{H} + \mathbf{R}.$$

The Kalman filter minimizes the error covariance of the estimated objects; therefore, the forecasts $\hat{\mathbf{x}}_{t+1|t}$ and $\hat{\mathbf{y}}_{t+1|t}$ are best estimators within the class of linear filters (i.e. forecasts that are linear functions of $(\mathbf{z}_t, \mathbf{\Omega}_{t-1})$). If we further assume that initial state $\mathbf{x}_{1|0}$ and innovations $\{\mathbf{e}_t, \mathbf{u}_t\}_{t=1}^{T}$ are multivariate Gaussian, then the forecasts are optimal among any functions of $(\mathbf{z}_t, \mathbf{\Omega}_{t-1})$.

11.4.1 Evaluation of the Likelihood Function using the Kalman Filter

One of the applications of the Kalman filter is the evaluation of unconditional likelihood for a DSGE model. Consider a state-space representation of the solution of the DSGE model:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{F}(\mu)\mathbf{x}_{t-1} + \mathbf{u}_t \\ \mathbf{u}_t &= \mathbf{G}(\mu)\mathbf{v}_t, \end{aligned}$$

where $\mathbf{x}_t$ is an $(r \times 1)$ vector of model variables, and $E(\mathbf{u}_t\mathbf{u}_t^{'}) = \mathbf{G}(\mu)E(\mathbf{v}_t\mathbf{v}_t^{'})\mathbf{G}(\mu)^{'} = \mathbf{Q}(\mu)$. A measurement equation maps $\mathbf{x}_t$ into the $n \times 1$ vector of observable variables $\mathbf{y}_t$:

$$\mathbf{y}_t = \mathbf{H}(\mu)^{'}\mathbf{x}_t + \mathbf{e}_t,$$

where $\mathbf{e}_t$ is an $n \times 1$ vector of measurement errors with $E(\mathbf{e}_t\mathbf{e}_j^{'}) = \mathbf{R}$ for $t = j$ and $\mathbf{0}$ otherwise. Given time-series data and the model's parameter values μ (so that $F(\mu)$, $G(\mu)$, $Q(\mu)$, and $H(\mu)$ are known), the Kalman filter infers a sequence of conditional distribution for $\mathbf{x}_t$ given $\mathbf{x}_{t-1}$ and evaluate the likelihood.

In order to implement the Kalman filter, assume that $\mathbf{e}_t$, $\mathbf{u}_t$, and $\mathbf{v}_t$ are normally distributed. The initial unconditional values are given by

$$\hat{\mathbf{x}}_{1|0} = \mathbf{0}, \quad \mathbf{P}_{1|0} = \mathbf{F}\mathbf{P}_{1|0}\mathbf{F}' + \mathbf{Q}$$

where $\text{vec}(\mathbf{P}_{1|0}) = (\mathbf{I} - \mathbf{F} \otimes \mathbf{F}')^{-1}\text{vec}(\mathbf{Q})$.

Given the initial values, the projection $\hat{\mathbf{x}}_{t|t-1}$ and its associated covariance matrix $\mathbf{P}_{t|t-1}$ are iterated on:

$$\begin{aligned} \hat{\mathbf{x}}_{t|t-1} &= \mathbf{F}\hat{\mathbf{x}}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{Q}, \end{aligned}$$

where $\text{vec}(\mathbf{P}_{t|t-1}) = (\mathbf{I}-\mathbf{F}\otimes\mathbf{F}')^{-1}\text{vec}(\mathbf{Q})$. These projections are then used to construct the conditional distribution of $\mathbf{y}_t$, $N(\hat{\mathbf{y}}_{t|t-1}, \mathbf{\Sigma}_{t|t-1})$, where the conditional mean $\hat{\mathbf{y}}_{t|t-1}$ and conditional variance matrix $\mathbf{\Sigma}_{t|t-1}$ are given by

$$\begin{aligned} \hat{\mathbf{y}}_{t|t-1} &= \mathbf{H}'\hat{\mathbf{x}}_{t|t-1} \\ \mathbf{\Sigma}_{t|t-1} &= E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'] \\ &= \mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R}. \end{aligned}$$

The likelihood function for $\mathbf{y}_t$ is thus given by:

$$L(\mathbf{y}_t|\boldsymbol{\mu}) = (2\pi)^{-m/2}|\boldsymbol{\Sigma}_{t|t-1}^{-1}|^{1/2}\exp\left[-\frac{1}{2}(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'\boldsymbol{\Sigma}_{t|t-1}^{-1}(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})\right].$$

The next iteration is initiated by updating $\hat{\mathbf{x}}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$:

$$\begin{aligned}\mathbf{x}_{t|t} &= \mathbf{x}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}\boldsymbol{\Sigma}_{t|t-1}^{-1}(\mathbf{y}_t - \mathbf{y}_{t|t-1}) \\ \mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}\boldsymbol{\Sigma}_{t|t-1}^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}.\end{aligned}$$

Finally, the likelihood from each iteration is multiplied to yield the sample likelihood:

$$L(\mathbf{y}|\boldsymbol{\mu}) = \prod_{t=1}^{T} L(\mathbf{y}_t|\boldsymbol{\mu}).$$

This likelihood function is maximized to yield the ML estimator of linearized DSGE models.

Appendix

11.A Examples of State-Space Representations

This appendix contains examples of the state-space representation for AR(p) and MA(p) processes. There are several ways of representing a given process in state-space form. For more examples, see Hamilton (1994, Ch. 13).

Example 1: Univariate AR(p) Process Consider a univariate AR(p) process:

$$y_{t+1} - \mu = \phi_1(y_t - \mu) + \phi_2(y_{t-1} - \mu) + \cdots + \phi_P(y_{t-p+1} - \mu) + \varepsilon_{t+1},$$

where $E(\varepsilon_t\varepsilon_j) = \sigma^2$ for $j \neq t$ and 0 otherwise. One example of the state-space representation for this process is

$$\mathbf{x}_t = \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix}, \mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \mathbf{u}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

$$\mathbf{y}_t = y_t,\ \mathbf{A}' = \mu,\ \mathbf{z}_t = 1,\ \mathbf{H}' = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}, \mathbf{e}_t = 0, \mathbf{R} = 0.$$

Example 2: Univariate MA(1) Process

For a univariate MA(1) process

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

where $E(\varepsilon_t\varepsilon_j) = \sigma^2$ for $j \neq t$ and 0 otherwise, the state-space representation is given by

$$\mathbf{x}_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \mathbf{u}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix}, \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\mathbf{y}_t = y_t,\ \mathbf{A}' = \mu,\ \mathbf{z}_t = 1,\ \mathbf{H}' = \begin{bmatrix} 1 & \theta \end{bmatrix}, \mathbf{e}_t = 0, \mathbf{R} = 0.$$

References

Altug, S., and R. A. Miller (1990): "Household Choices in Equilibrium," *Econometrica*, 58, 543–570.

Ambler, S., E. Cardia, and C. Zimmermann (2004): "International Business Cycles: What are the Facts?," *Journal of Monetary Economics*, 51(2), 257–276.

Atkeson, A., and M. Ogaki (1996): "Wealth-Varying Intertemporal Elasticities of Substitution: Evidence from Panel and Aggregate Data," *Journal of Monetary Economics*, 38, 507–534.

Calvo, G. A. (1983): "Staggered Prices and in a Utility-Maximizing Framework," *Journal of Monetary Economics*, 12(3), 383–398.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113(1), 1–45.

Eichenbaum, M., and J. D. M. Fisher (2007): "Estimating the Frequency of Price Re-Optimization in Calvo-Style Models," *Journal of Monetary Economics*, 54(7), 2032–2047.

Galí, J., and M. Gertler (1999): "Inflation Dynamics: A Structural Econometric Analysis," *Journal of Monetary Economics*, 44(2), 195–222.

García-Cicco, J., R. Pancrazi, and M. Uribe (2009): "Real Business Cycles in Emerging Countries? Expanded Version," Manuscript.

Hamilton, J. D. (1994): *Time Series Analysis*. Princeton University Press, Princeton.

Kalman, R. E. (1960): "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering*, 82(1), 35–45.

Kim, H., and M. Ogaki (2009): "Purchasing Power Parity and the Taylor Rule," Working Paper No. 09-03, Department of Economics, Ohio State University.

Kleibergen, F., and S. Mavroeidis (2009): "Weak Instrument Robust Tests in GMM and the New Keynesian Phillips Curve," *Journal of Business and Economic Statistics*, 27(3), 293–311.

Magnusson, L. M., and S. Mavroeidis (2009): "Identification-Robust Minimum Distance Estimation of the New Keynesian Phillips Curve," Working Paper 0904, Department of Economics, Tulane University.

Sbordone, A. M. (2002): "Prices and Unit Labor Costs: A New Test of Price Stickiness," *Journal of Monetary Economics*, 49(2), 265–292.

——— (2005): "Do Expected Future Marginal Costs Drive Inflation Dynamics?," *Journal of Monetary Economics*, 52(6), 1183–1197.