Introduction to Structural Estimation of DSGE Models*

David Murakami†

9th May 2021

Contents

*The references for GMM, SMM, and ML are Davidson and MacKinnon (2004), Hayashi (2000), and Miao (2020); I use Ljungqvist and Sargent (2018) for the Kalman filter and its examples. Additionally, I follow lecture notes by Wouter Den Haan, James Duffy, Petr Sedláček, and Karl Whelan.

†St Cross College, University of Oxford. Email: david.murakami@economics.ox.ac.uk

1 Introduction

Most economic data are generated using economic decision rules and state processes. There are three ways to determine from the data what parameter values an economic agent used in the decision rules

and state processes: calibration, reduced-form estimation, and structural estimation. Using reduced-form estimation methods, the parameters can be estimated directly by specifying functional forms for decision rules or state-transition equations, independent of behavioural theory – this will not be our focus here.

These notes will cover calibration and structural estimation methods: generalised method-of-moments (GMM) and the simulated method-of-moments (SMM). We will be focusing on the concepts of DGSE model estimation, and so we will also cover maximum likelihood (ML) and the Kalman filter.

2 Calibration

The main idea of calibration is that you select model parameters ("pin them down") by selected real-world features. Calibration became mainstream in macroeconomics following the seminal paper by Kydland and Prescott (1982). You choose parameter values on the basis of microeconomic evidence and then to compare the model's predictions concerning the variances and covariances (referred to as "moments" or "business-cycle moments") of various series with those in the data. Romer (2012) gives a fairly concise summary of calibration in the context of the real-business-cycle (RBC) model.

Calibration has two potential advantages over estimating models econometrically. First, because parameter values are selected on the basis of microeconomic evidence, a large body of information beyond that usually employed can be brought to bear, and the models can therefore be held to a higher standard. Second, the economic importance of a statistical rejection, or lack of rejection, of a model is often hard to interpret. A model that fits the data well along every dimension except one unimportant one may be overwhelmingly rejected statistically. Or a model may fail to be rejected simply because the data are consistent with a wide range of possibilities.

To see how calibration works in practice, consider the baseline RBC model of Hansen (1985) and Prescott (1986). Note that this model does not feature government, and the trend component of technology is not assumed to follow a simple linear path; instead, a smooth but nonlinear trend is removed from the data before the model's predictions and actual fluctuations are compared.[1]

We consider the parameter values proposed by Hansen and Wright (1992). Based on data on factor shares, the capital-output ratio, and the investment-output ratio, Hansen and Wright set $\alpha = 0.36$, $\delta = 0.025$ per quarter, and $\beta = 0.99$ per quarter. Based on the average division of discretionary time between work and non-work activities, they set b (the coefficient for the disutility of labour) to 2. They choose parameters of the process for technology on the basis of the empirical behaviour of the Solow residual,[2]

$$R_t \equiv \Delta \ln Y_t - [\alpha \Delta \ln K_t + (1-\alpha)\Delta \ln L_t] .$$

[1] The data is detrended using a Hodrick-Prescott (HP) filter.

[2] The Solow residual is a measure of all influences on output growth other than the contributions of capital and labour through private marginal products.

Under the assumptions of RBC theory, the only such other influence on output is technology, and so the Solow residual is a measure of technological change. Based on the behaviour of the Solow residual, Hansen and Wright set $\rho_A = 0.95$ and the standard deviation of the quarterly ϵ_A's to 1.1 percent.[3]

Table 1: Calibrated RBC Model verses Actual Data

	US Data (1947-1991)	Baseline RBC Models
σ_Y	1.92	1.30
σ_C/σ_Y	0.45	0.31
σ_I/σ_Y	2.78	3.15
σ_L/σ_Y	0.96	0.49
$\text{Corr}(L, Y/L)$	-0.14	0.93

Table 1 shows the model's implications for some key features of fluctuations. The figures in the first column are from actual US data; those in the second column are from the model. All of the numbers are based on the deviation-from-trend components of the variables, with the trends found using the non-linear procedure employed by Prescott and Hansen.

The first line of the table reports the standard deviation of output, σ_Y. The model produces output fluctuations that are only moderately smaller than those observed in practice. This finding is the basis for Prescott's famous conclusion that aggregate fluctuations are not just consistent with a competitive, neoclassical model, but are predicted by such a model. The second and third rows of the table show that both in the US and in the model, consumption is considerably less volatile than output, and investment is considerably more volatile.

The final two lines of the table show that the baseline RBC model is less successful in its predictions about the contributions of variations in labour input and in output per unit of labour input to aggregate fluctuations. In the US economy, labour input is nearly as volatile as output; in the model it is much less so. And in the US, labour input and productivity are essentially uncorrelated; in the model they move together closely.

Thus, a simple calibration exercise can be used to identify a model's major success and failures. In doing so, it suggests ways in which the model might be modified to improve its fit with the data. For example, additional sources of shocks would be likely to increase output fluctuations and to reduce the correlation between movements in labour input and in productivity. Indeed, Hansen and Wright show that, for their suggested parameter values, adding government-purchases shocks lowers the correlation of L and Y/L from 0.93 to 0.49; the change has little effect on the magnitude of output fluctuations,

[3] In addition, Prescott argues that, under the assumption that technology multiplies an expression of the form $F(K, L)$, the absence of a strong trend in capital's share suggests that $F(\cdot)$ is approximately Cobb-Douglas. Similarly, he argues on the basis of the lack of a trend in leisure per person and of studies of substitution between consumption in different periods that

$$u_t = \ln c_t + b(1 - l_t)$$

provides a good approximation to the instantaneous utility function. Thus, the choices of functional form are not arbitrary.

however.

Of course, calibration has disadvantages as well. DSGE models have long since moved away from the highly Walrasian nature of RBC models. As a result, calibration exercises no longer rely on the original idea of using microeconomic evidence to tie down essentially all the relevant parameters and functional forms: given the models' wide variety of features, they have some flexibility in matching the data. As a result, we do not know how informative it is when they match important moments of the data relatively well. Nor, because the models are generally not tested against alternatives, do we know whether there are other, perhaps completely different, models that can match the moments just as well.

Further, given the state of economic knowledge, it is not clear that matching the major moments of the data should be viewed as a desirable feature of a model. Even the most complicated models of fluctuations are grossly simplified descriptions of reality. It would be remarkable if none of the simplifications had qualitatively important effects on the models' implications. But given this, it is hard to determine how informative the fact that a model does or does not match aggregate data is about its overall usefulness.

It would be a mistake to think that the only alternative to calibration is formal estimation of fully specified models. Often, the alternative is to focus more narrowly.[4] Researchers frequently assess models by considering the microeconomic evidence about the reasonableness of the models' central building blocks or by examining the models' consistent with a handful of "stylised facts" that the modellers view as crucial.

Unfortunately, there is little evidence concerning the relative merits of different approaches to evaluating macroeconomic models – there is no ideal "cookbook" method to this. Researchers use various mixes and types of calibration exercises, formal estimation, examination of the plausibility of the ingredients, and consideration of consistency with specific facts. At this point, choices among these approaches seem to be based more on researchers' "tastes" than on a body of knowledge about the strengths and weaknesses of the approaches.

3 Generalised Method of Moments

The generalised method of moments (GMM), formalised by Hansen (1982), is an estimation method that exploits the sample moment counterparts of population moment conditions (orthogonality conditions) of the data-generating process.

The idea of matching moments is similar to calibration: you parameterise by a set of moments (features) of the data, and then you judge the model performance by a different set of moments. Matching moments adds statistical rigour, as you estimate based on limited information and you can

[4] This was a key point that Toni Braun and Fumio Hayashi stressed in their seminars and lectures.

conduct hypothesis testing. We won't go over all the background of GMM and SMM,[5] but we will go through a brief overview of GMM.

3.1 Review of method of moments estimation

Consider the χ^2 distribution. Recall that if the random vector, $\mathbf{z}$, is such that its components $z_1, ..., z_k$ are mutually independent standard normal random variables. An easy way to express this is to write $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, the random variable

$$X \equiv \|\mathbf{z}\|^2 = \mathbf{z}^\top \mathbf{z} = \sum_{i=1}^{k} z_i^2,$$

is said to follow the χ^2 distribution with k degrees of freedom. Suppose that we have an IID sample $(x_1, ..., x_T)$, and are interested in estimating k. How do we go about doing this?

A very natural way to estimate parameters – and, as an alternative to ML estimation – is to replace population means by sample means. This technique is called the method of moments (MM), and it is one of the most widely-used estimation methods in statistics. As the name implies, it can be used with moments other than the mean. In general, MM estimates population moments by the corresponding sample moments. In order to apply this method, we must use the facts that population moments are expectations, and that regression models are specified in terms of the conditional expectations of the error terms.

Here, our first two moments are

$$\begin{aligned} m_1(k) &\equiv \mathbb{E}[X] = k, \\ m_2(k) &\equiv \mathbb{E}[X^2] = k(k+2). \end{aligned}$$

The sample estimate of the first moment is given by

$$\frac{1}{T} \sum_{t=1}^{T} x_t,$$

so the MM estimator of k using the first moment is

$$\frac{1}{T} \sum_{t=1}^{T} x_t = \hat{k}.$$

The moment condition is

$$m_1(k) \equiv \mathbb{E}[X] = k,$$

[5] Separate notes based on Davidson and MacKinnon (2004) and Hayashi (2000) are available for this.

and the sample equivalent of the moment condition is

$$\hat{m}_1 = \frac{1}{T}\sum_{t=1}^{T} x_t = \hat{k}.$$

We write the MM estimator so that the moment condition is equal to zero. The population moment condition is:

$$m_1(k) \equiv \mathbb{E}[z] - k = 0,$$

and the sample moment condition is:

$$\hat{m}_1(\hat{k}) \equiv \frac{1}{T}\sum_{t=1}^{T} z_t - \hat{k} = 0.$$

We could've also used the second moment to estimate m:

$$m_2(k) \equiv \mathbb{E}[X^2] = k(k+2),$$

where the sample equivalent of the second moment is

$$\hat{m}_2 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t^2,$$

so the MM estimator of k solves the following moment condition

$$\hat{m}_2 \equiv \frac{1}{T}\sum_{t=1}^{T} x_t^2 - \hat{k}(\hat{k}+2) = 0.$$

3.1.1 Example: Estimating the simple linear regression model

Let us review MM estimation for the simple linear regression model,

$$y_t = \beta_1 + \beta_2 x_t + u_t.$$

The error term for observation t is

$$u_t = y_t - \beta_1 - \beta_2 x_t,$$

and, according to classical assumptions, the expectation of this error term is zero. Since we have T error terms for a sample size of T, we can consider the sample mean of the error terms:

$$\frac{1}{T}\sum_{t=1}^{T} u_t = \frac{1}{T}\sum_{t=1}^{T}(y_t - \beta_1 - \beta_2 x_t). \tag{1}$$

We would like to set this sample mean equal to zero.

Suppose to begin with, we assume $\beta_2 = 0$. This reduces the number of parameters in the model to just one. In that case, there is just one value of β_1 which allows the RHS of (1) to equal zero. The equation defining this value is

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \beta_1) = 0.$$

Since β_1 is common to all the observations and thus does not depend on the index t, we can write this as

$$\frac{1}{T}\sum_{t=1}^{T} y_t - \beta_1 = 0,$$

which gives us the estimate of β_1:

$$\hat{\beta}_1 = \frac{1}{T}\sum_{t=1}^{T} y_t.$$

Thus, if we wish to estimate the population mean of the y_t, which is what β_1 is in our model when $\beta_2 = 0$, MM estimation tells us to use the sample mean as our estimate.

Now, put β_2 back into our model:

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - \beta_1 - \beta_2 x_t) = 0. \tag{2}$$

But this is one equation with two unknowns. In order to obtain another equation, we can use the fact that our model specifies that the mean of u_t is 0 conditional on the explanatory variable, x_t. The conditional mean assumption implies that not only is $\mathbb{E}[u_t] = 0$ (through the Law of Iterated Expectations (LIE)), but that $\mathbb{E}[x_t u_t] = 0$ as well:

$$\mathbb{E}[x_t u_t] = \mathbb{E}[\mathbb{E}[x_t u_t | x_t]] = \mathbb{E}[x_t \mathbb{E}[u_t | x_t]] = 0.$$

Thus, we can supplement (2) by the following equation, which replaces the population mean above by the corresponding sample mean,

$$\frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \beta_1 - \beta_2 x_t) = 0. \tag{3}$$

Thus, we have two equations, (2) and (3), with two unknowns, β_1 and β_2.

Since β_1 and β_2 do not depend on t, these two equations can be written as

$$\begin{aligned}\beta_1 + \left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\beta_2 &= \frac{1}{T}\sum_{t=1}^{T} y_t,\\ \left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\beta_1 + \left(\frac{1}{T}\sum_{t=1}^{T} x_t^2\right)\beta_2 &= \frac{1}{T}\sum_{t=1}^{T} x_t y_t,\end{aligned}$$

or in matrix form as:

$$\begin{bmatrix} T & \sum_{t=1}^{T} x_t \\ \sum_{t=1}^{T} x_t & \sum_{t=1}^{T} x_t^2 \end{bmatrix}\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} = \begin{bmatrix}\sum_{t=1}^{T} y_t \\ \sum_{t=1}^{T} x_t y_t\end{bmatrix}. \tag{4}$$

Equations (4) can be rewritten much more compactly (given our moment conditions):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}.$$

Thus, it is clear that we can rewrite those equations as

$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{y},$$

which, of course, yields the famous OLS estimator

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}.$$

Thus, it should be quite obvious that the OLS estimator is a special case of the MM estimator.

3.2 General moment conditions

OLS is a special case of MM, and moment estimators are a much more general class of estimators than OLS. The general formulation is that if $\mathbf{z}_t$ is a vector stationary data, such as

$$\begin{aligned}\mathbf{z}_t &= (x_t),\\ \mathbf{z}_t &= (x_t, y_t)^\top,\end{aligned}$$

and we let $\boldsymbol{\theta}$ denote the vector of parameters to be estimated, then the moment conditions can be written as

$$\mathbf{g}(\boldsymbol{\theta}) \equiv \mathbb{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{z}_t)] = \mathbf{0}.$$

Theorem 1 (Properties of the MM Estimator)**.** *Let $\mathbf{h}(\boldsymbol{\theta}, \mathbf{z}_t)$ be a $k \times 1$ vector of moment conditions, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters and $\mathbf{z}_t$ is a sequence of stationary data.*

The MM estimator, $\hat{\boldsymbol{\theta}}$, sets the sample moment conditions to zero:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}(\hat{\boldsymbol{\theta}},\mathbf{z}_t)=\mathbf{0}.$$

Under some mild technical conditions, the MM estimator has the following properties:

1) The estimator is consistent: $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$.

2) The estimator is asymptotically normal and root-n consistent:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta})\xrightarrow{d}\mathcal{N}(\mathbf{0},\mathbf{V}).$$

In the $\chi^2(k)$ example, with the first moment we had $g_1(k,x_t)=x_t-k$, and with the second moment we had $g_2(k,x_t)=x_t^2-k(k+2)$. In the univariate OLS case we have $h(\beta,(x_t,y_t))=x_tu_t=x_t(y_t-x_t\beta)$. The MM estimator sets the sample mean of the moment conditions to 0, which we can write as the following for the scalar case:

$$g(\hat{\theta})=\frac{1}{T}\sum_{t=1}^{T}h(\hat{\theta},\mathbf{z}_t)=0.$$

Note that when the LLN applies, we can replace the population moment by the empirical moment.

In the multivariate case, where $\mathbf{X}_t$ is a $1\times k$ row vector and $\boldsymbol{\beta}$ is a $k\times 1$ vector,

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u},$$

where we now have k moment conditions:

$$\mathbb{E}\left[\mathbf{h}(\boldsymbol{\beta};(\mathbf{X}_t,y_t)\right]=\mathbb{E}[\mathbf{X}_t^\top u_t]=\mathbf{0},$$

where have k moment conditions for k parameters.

The sample analog is:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{X}_t^\top(y_i-\mathbf{X}_t\hat{\boldsymbol{\beta}})=0,$$

$$\hat{\boldsymbol{\beta}}_{\text{MM}}=\left(\sum_{t=1}^{T}\mathbf{X}_t^\top\mathbf{X}_t\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{X}_t^\top y_t\right).$$

Recall that if the assumption of $\mathbb{E}[\mathbf{X}_t^\top u_t]=\mathbf{0}$ is violated, OLS is inconsistent and biased. In such a case we can use an instruments, say $\mathbf{Z}_t$, that are correlated with $\mathbf{X}_t$ but uncorrelated with u_t. The MM moment condition is thus

$$\mathbb{E}[\mathbf{Z}_t^\top u_t]=\mathbf{0},$$

and the sample analog becomes

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}_t^\top(y_t - \mathbf{X}_t\hat{\boldsymbol{\beta}}) = 0,$$

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \left(\sum_{t=1}^{T}\mathbf{Z}_t^\top\mathbf{X}_t\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{Z}_t^\top y_t\right),$$

which is equal to the instrumental variable (IV) regressor.

3.2.1 MLE as GMM

Consider the log-likelihood function:

$$l = \frac{1}{n}\sum_{i=1}^{n}\log f(y_i|x_i;\boldsymbol{\theta}),$$

and the population expectation of the FOC:

$$\mathbb{E}\left[\frac{\partial l}{\partial\theta_k}\right] = 0.$$

The GMM sample equivalent is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\log f(y_i|x_i;\boldsymbol{\theta})}{\partial\theta_k} = 0,$$

and thus we have k nonlinear equations with k unknowns.

3.2.2 MM two-stage estimation with measurement error

Some factor models need to be estimated in two steps. Suppose we have returns of n different stocks. First, estimate the "beta" of each stock i:

$$R_{i,t} = \beta_i R_{M,t} + e_{i,t}.$$

Then, estimate the "factor risk premium" with $\hat{\beta}_i$ as an explanatory variable:

$$\mu_i = \lambda\hat{\beta}_i + u_i,$$

where $\mu_i = \mathbb{E}[R_i]$.

The problem here is that the first stage estimation of $\hat{\beta}_i$ has estimation error that needs to be accounted for in the second stage estimation. The solution to this is to stack both equations together

and estimate $\boldsymbol{\theta} = (\beta, \lambda)$ simultaneously:

$$\mathbf{g}(\boldsymbol{\theta}) = \mathbb{E}\begin{bmatrix} e_t R_{M,t} \\ u_i \beta_i \end{bmatrix} = \mathbb{E}\begin{bmatrix} (R_{i,t} - \beta R_{M,t}) R_{M,t} \\ (\mu_i - \beta_i \lambda)\beta_i \end{bmatrix} = \mathbf{0}.$$

We will look again at this later, but this GMM system takes the effect of estimation uncertainty of $\hat{\beta}_i$ on the standard error of $\hat{\lambda}$ correctly into account. GMM is useful for many multi-step estimation methods.

3.3 From MM to GMM

So far in the simple examples considered, the number of moment conditions equal the number of parameters – where we say that the MM is exactly identified. If there are more parameters than moment conditions, then we say that the MM is under-identified and the parameters cannot be estimated. But if the number of moment conditions exceeds the number of parameters, then we say that the MM is overidentified, which is where we use GMM.

GMM is a versatile estimation method. It is consistent and asymptotically normal under mild assumptions, but it does require a bit of a trick to use – we essentially have to convert a given problem into a set of moment conditions. Once an estimator is written as a moment condition, we know it is consistent and asymptotically normal. It's not all roses, however: GMM can have poor small sample properties – but then again, what estimation procedure doesn't suffer from this.

The recommended texts are, as mentioned before, the wonderful texts by Davidson and MacKinnon (2004) and Hayashi (2000). Hayashi probably provides the most in-depth treatment, and anyone comfortable with least squares estimation and asymptotic theory will feel right at home; Davidson and MacKinnon target those who are more familiar with orthogonal projection matrices. Either way, they're both good books. Cochrane (2005) is also good when it comes to framing asset pricing into the GMM framework. We won't go as deep as those books, but we'll go over a brief overview.

Consider our previous example where $X \sim \chi^2(k)$. The first 2 moments were

$$\begin{aligned} m_1(k) &\equiv \mathbb{E}[X] = k, \\ m_2(k) &\equiv \mathbb{E}[X^2] = k(k+2). \end{aligned}$$

In MM, we used either the first or second moment to estimate k. In GMM, we combine moment conditions to estimate k. In MM, we have have as many moment conditions as parameters. hence, we can pick the parameters to set the moment conditions exactly to zero. If we have more moment conditions than parameters, not all moment conditions can be exactly satisfied. With GMM, we pick the parameters that minimise a weighted average of the moment conditions.

In the $\chi^2(k)$ example, the moment conditions were

$$\begin{aligned} m_1(k) &\equiv \mathbb{E}[X] - k = 0, \\ m_2(k) &\equiv \mathbb{E}[X^2] - k(k+2) = 0. \end{aligned}$$

Let $\mathbf{m}(k) = (m_1(k), m_2(k))^\top$ be the vector of moment conditions and let $\mathbf{W}$ be a symmetric positive definite matrix:

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix}.$$

The GMM estimator of k minimises the quadratic form

$$\begin{aligned} \hat{m}_{\text{GMM}} &= \arg\min_k \mathbf{m}(k)^\top \mathbf{W} \mathbf{m}(k), \\ \mathbf{m}(k)^\top \mathbf{W} \mathbf{m}(k) &= w_{11} m_1(k)^2 + 2 w_{12} m_1(k) m_2(k) + w_{22} m_2(k)^2. \end{aligned}$$

Thus, we have the following population moments:

$$\begin{aligned} m_1(k) &\equiv \mathbb{E}[X] - k = 0, \\ m_2(k) &\equiv \mathbb{E}[X^2] - k(k+2) = 0, \end{aligned}$$

and the sample equivalents:

$$\begin{aligned} \hat{m}_1(k) &\equiv \frac{1}{T} \sum_{t=1}^{T} x_t - k = 0, \\ \hat{m}_2(k) &\equiv \frac{1}{T} \sum_{t=1}^{T} x_t^2 - k(k+2) = 0. \end{aligned}$$

Suppose we had the given sample:

$$\begin{aligned} \hat{m}_1(k) &\equiv \frac{1}{T} \sum_{t=1}^{T} x_t = 9.47, \\ \hat{m}_2(k) &\equiv \frac{1}{T} \sum_{t=1}^{T} x_t^2 - k(k+2) = 104.18. \end{aligned}$$

The MM estimator for $\mathbb{E}[X] - k = 0 \implies \hat{k} = 9.47$, which is a special case of GMM with

$$\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

The MM estimator for $\mathbb{E}[X^2] - k(k+2) = 0 \implies \hat{k} = 9.26$, which is a special case of GMM with

$$\mathbf{W} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

We could also have something like

$$\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 1/10 \end{bmatrix},$$

which would lead to $\hat{k} = 9.30$.

So the question for us is: how do we pick the weighting matrix, $\mathbf{W}$? Unfortunately, there's no hard rule for this. It comes down to intuition. You put more weight on moments that are "more informative" about the true $\boldsymbol{\theta}$. This is akin to how we choose appropriate priors in Bayesian estimation.

GMM is a very powerful way of looking at an estimation problem. All we need is a moment condition that holds.

3.4 The GMM methodology

Following Hansen (1982), suppose you have an economic model which implies a set of r moment conditions that take the form

$$\mathbb{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t)] = \mathbf{0},$$

where $\mathbf{Z}_t$ is a $1 \times l$ vector of variables known at time t, and $\boldsymbol{\theta}$ is a $k \times 1$ vector of coefficients we seek to estimate. The above is a population mean. The sample equivalent is

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t) \equiv \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t).$$

The GMM estimator of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ that minimises the scalar[6]

$$Q(\boldsymbol{\theta}, \mathbf{Z}_t) = \mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t)^\top \mathbf{W} \mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t), \tag{5}$$

where $\mathbf{W}$ is a $r \times r$ positive definite weighting matrix.

If $r = k$, then the number of parameters to be estimated is equal to the number of moment conditions. Then typically the objective function (5) will be minimised by setting

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MM}}, \mathbf{Z}_t) = \mathbf{0}.$$

[6]Technically, the weighting matrix should be denoted as something like $\mathbf{W}_T$, which is positive semidefinite and converges in probability to the positive definite matrix $\mathbf{W}$.

If $r > k$, we cannot set all moment conditions exactly to zero. Instead, we have

$$\hat{\boldsymbol{\theta}}_{\text{GMM}} = \arg\min \left\{ \mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t)^\top \mathbf{W} \mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t) \right\}.$$

The quadratic form can be minimised with respect to $\boldsymbol{\theta}$ using analytic or numerical methods.

Theorem 2 (Hansen (1982))**.** *If $\mathbf{Z}_t$ are strictly stationary*

$$\mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t) \xrightarrow{p} \mathbb{E}\left[\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t)\right].$$

If $\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t)$ is continuous in $\boldsymbol{\theta}$, then the GMM estimator is consistent:

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0.$$

GMM is consistent for any positive semidefinite weighting matrix.

But how should we choose the weighting matrix? And what are the asymptotic properties of the GMM estimator?

3.4.1 Example: Consumption CAPM

Consider the household stochastic discount factor, $M_{t,t+1}$, that prices an asset j with payoff $X_{j,t+1}$:

$$\begin{aligned} P_{j,t} &= \mathbb{E}_t[M_{t,t+1} X_{j,t+1}], \\ R_{j,t+1} &= \frac{X_{j,t+1}}{P_{j,t}} = \frac{P_{j,t+1} + D_{j,t+1}}{P_{j,t}}, \\ \implies \mathbb{E}[M_{t,t+1} R_{j,t+1}] &= 1. \end{aligned}$$

Recall that the idea from consumption CAPM was that $M_{t,t+1}$ depends on consumption, C_t, and preferences, $U(C_t)$:

$$\mathbb{E}_t\left[\beta \frac{U'(C_{t+1})}{U'(C_t)} R_{j,t+1}\right] = 1,$$

where we assume that the household is risk averse and has a well behaved utility function (e.g. CRRA-form):

$$\begin{aligned} U(C_t) &= \beta^t \frac{C_t^{1-\gamma}}{1-\gamma}, \\ U'(C_t) &= \beta^t C_t^{-\gamma}, \\ \implies \mathbb{E}_t\left[\beta \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R_{j,t+1} - 1\right] &= 0. \end{aligned}$$

But notice that the last equation is a moment condition!

We can test whether this equation holds in the data. We have returns $R_{j,t+1}$ of the $j = 1, ..., J$ assets, aggregate consumption data for C_t, and the following:

$$\begin{aligned}
\mathbb{E}[\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t)] &= \mathbf{0}, \\
\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t) &= (h_1(\boldsymbol{\theta}, \mathbf{Z}_t), ..., h_J(\boldsymbol{\theta}, \mathbf{Z}_t))^\top, \\
h_j(\boldsymbol{\theta}, \mathbf{Z}_t) &= \beta \left(\frac{C_{t+1}}{C_t} \right)^{-\gamma} R_{j,t+1} - 1, \\
\boldsymbol{\theta} &= (\beta, \gamma)^\top, \\
\mathbf{Z}_t &= (R_{1,t}, ..., R_{J,t}, C_t)^\top.
\end{aligned}$$

There are J moment conditions (one for each asset) and 2 parameters. So, we have to pick a weighting matrix.

One option is $\mathbf{W} = \mathbf{I}$ – i.e., all moments have the same weight. But perhaps a better idea is to follow what we do with GLS: observations are weighted according to their variance, so that we put more weight on moments whose variance is smaller. If the data is IID:

$$\begin{aligned}
\mathbf{W} &= \mathbf{S}^{-1}, \\
\mathbf{S} &= \mathbb{E}\left[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t) \mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)^\top \right] \\
&= \text{Var}\left(\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t) \right).
\end{aligned}$$

The sample equivalent is

$$\begin{aligned}
\hat{\mathbf{W}} &= \hat{\mathbf{S}}^{-1}, \\
\hat{\mathbf{S}} &= \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t) \mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)^\top.
\end{aligned}$$

3.4.2 Example: The χ^2 distribution

Recall that if the random vector, $\mathbf{x}$, is such that its components $x_1, ..., x_k$ are mutually independent standard normal random variables. An easy way to express this is to write $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then, the random variable

$$z = \sum_{i=1}^{k} x_i^2 \sim \chi^2(k).$$

The first four moments are

$$\begin{aligned}
m_1(k) &= k, \\
m_2(k) &= k(k+2), \\
m_3(k) &= k(k+2)(k+4), \\
m_4(k) &= k(k+2)(k+4)(k+6).
\end{aligned}$$

Suppose we have a sample of $\mathbf{z} = (z_1, ..., z_T)^\top$ observations. The moment conditions for the first two moments are

$$\begin{aligned}
g_1(k, \mathbf{z}) &\equiv m_1(k) - k = 0, \\
g_2(k, \mathbf{z}) &\equiv m_1(k) - k(k+2) = 0.
\end{aligned}$$

The GMM estimator using the first 2 moments and $\mathbf{W} = \mathbf{I}$ minimises

$$\begin{bmatrix} \hat{m}_1(\mathbf{z}) - k \\ \hat{m}_2(\mathbf{z}) - k(k+2) \end{bmatrix}^\top \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{m}_1(\mathbf{z}) - k \\ \hat{m}_2(\mathbf{z}) - k(k+2) \end{bmatrix}.$$

Next, let's compute GMM using the optimal $\mathbf{W} = \mathbf{S}^{-1}$:

$$\mathbf{S} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \mathbb{E} \begin{bmatrix} y_t - m_1(k) \\ y_t^2 - m_2(k) \end{bmatrix} \begin{bmatrix} y_t - m_1(k) \\ y_t^2 - m_2(k) \end{bmatrix}^\top,$$

where we have

$$\begin{aligned}
S_{11} &= \mathbb{E}y_t^2 - 2m_1^2 + m_1^2 = k(k+2) - 2 = 2k, \\
S_{12} &= m_3(k) - m_1(k)m_2(k) = k(k+2)(k+4) - k^2(k+2) = 4k(k+2), \\
S_{22} &= m_4(k) - [m_2(k)]^2 = k(k+2)(k+4)(k+6) - k^2(k+2)^2 = 8k(k+2)(k+3).
\end{aligned}$$

Recall our earlier example:

$$\begin{aligned}
\hat{m}_1(k) &\equiv \frac{1}{T}\sum_{t=1}^{T} x_t = 9.47, \\
\hat{m}_2(k) &\equiv \frac{1}{T}\sum_{t=1}^{T} x_t^2 = 104.18.
\end{aligned}$$

For $\mathbf{W} = \mathbf{I}$, $\hat{k} = 9.26$, and so if we compute $\mathbf{S}$ for $k = 9.26$:

$$\mathbf{S} = \begin{bmatrix} 18.51 & 416.78 \\ 416.78 & 10216.43 \end{bmatrix},$$
$$\mathbf{W} = \mathbf{S}^{-1} = \begin{bmatrix} 0.66 & -0.03 \\ -0.03 & 0.01 \end{bmatrix}.$$

Thus, $\mathbf{W} = \mathbf{I}$ and $\mathbf{W} = \mathbf{S}^{-1}$ are very different. Optimal GMM puts almost all the weight on the first moment.[7]

In practice, $\mathbf{S}$ cannot be computed analytically and has to be estimated. Many different estimators for $\mathbf{S}$ have been proposed. The most popular one is the Newey-West (1987) estimator:

$$\hat{\mathbf{S}} = \sum_{j=-q}^{q} \left(\frac{q - |j|}{q} \right) \frac{1}{T} \sum_{t=q+1}^{T-q} \left[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{z}_t) \right] \left[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{z}_{t-j}) \right]^{\top},$$

which down-weights higher-order autocorrelations, only autocorrelations up to lag q are used, q must be chosen ex-ante, and $\hat{\mathbf{S}}$ depends on $\hat{\boldsymbol{\theta}}$ which in turn depends on $\hat{\mathbf{S}}$. Thus the procedure to estimate $\mathbf{S}$ is iterative:

1. Obtain an initial estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}^{(1)}$, by minimising $Q(\boldsymbol{\theta}, \mathbf{z})$ for a given weighting matrix, usually $\mathbf{W} = \mathbf{I}$.
2. Use the initial estimate $\hat{\boldsymbol{\theta}}^{(1)}$ to produce an initial estimate of $\hat{\mathbf{S}}^{(1)}$.
3. Re-minimise $Q(\boldsymbol{\theta}, \mathbf{z})$ using the initial estimates $\hat{\mathbf{S}}^{(1)}$ to arrive at a new estimate, $\hat{\boldsymbol{\theta}}^{(2)}$.
4. One can continue iterating in this manner until estimates at successive iterations converge. In practice, usually one stops at $\hat{\boldsymbol{\theta}}^{(2)}$.

Note that $\hat{\mathbf{S}}$ is often close to being singular. Research has shown that computing $\hat{\mathbf{S}}^{-1}$ is often numerically unstable. The reason for this is that inverting large matrices is computationally difficult if they are close to being singular.

In practice: If the units of all moments are comparable, then the most robust results are obtained with $\mathbf{W} = \mathbf{I}$. If the units are different, then redefine the moments. The general rule of thumb is to first try the identity matrix as the optimal weighting matrix before attempting to set $\mathbf{W} = \hat{\mathbf{S}}$. If the results are substantially different figure out why. Is $\hat{\mathbf{S}}^{-1}$ poorly behaved? Is OLS very inefficient? Think about similar tradeoffs when considering OLS vs GLS.

[7] Note, of course, the optimal weighting matrix depends on k, which we don't know in practice.

3.5 Asymptotic properties of GMM

Theorem 3 (Hansen (1982))**.** *Assume that the stochastic process that generates the data* $\mathbf{Z}$ *is ergodic and stationary. Under certian regularity conditions, the GMM estimator is asymptotically normal:*

$$\sqrt{n}\mathbf{g}(\boldsymbol{\theta}, \mathbf{Z}_t) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{S}),$$
$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1}\right),$$

where

$$\mathbf{D} = \left[\frac{\partial g_1(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)}{\partial \boldsymbol{\theta}^\top} \quad \dots \quad \frac{\partial g_r(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)}{\partial \boldsymbol{\theta}^\top}\right].$$

The asymptotic normality of the GMM estimator is an important result. If an estimator can be written as a moment condition, it is generally consistent and asymptotically normal. But the crucial assumption is that the data is stationary.

Many standard problems can be written in GMM form. The real power of GMM is that one framework can handle a lot of interesting problems. Usually the moment conditions are directly implied by the definition of the estimator.

3.5.1 Example: OLS as a GMM estimator

We have our usual assumptions, starting with the data generating process (DGP):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

with the moment conditions,

$$\mathbf{g}(\boldsymbol{\beta}, \mathbf{Z}_i) = \mathbb{E}[\mathbf{h}(\mathbf{Z}_i, \boldsymbol{\beta})] = \mathbb{E}[\mathbf{X}_i^\top u_i] = \mathbb{E}[\mathbf{X}_i^\top (y_i - \mathbf{X}_i\boldsymbol{\beta})] = \mathbf{0},$$

where $\mathbf{Z}_t = (\mathbf{X}_t, y_t)$. Note that $\mathbf{X}_t$ is $1 \times k$, so there are k moment conditions for k parameters, hence GMM is exactly identified in this case. The sample analog is:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^\top (y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) = 0,$$
$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{X}_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^\top y_i\right),$$

and the GMM estimator is identical to the OLS estimator.

The GMM estimator is also asymptotically normal

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1}\right).$$

How about the variance-covariance matrix $(\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1}$?

$$\mathbf{D}^\top = \frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}^\top} = -\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top \mathbf{X}_i\right) = -\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{X}_i\right].$$

For simplicity, let's assume that $\mathbf{X}_i$ is a scalar, x_i, and impose the OLS assumptions of homoskedasticity and no serial correlation:

$$\begin{aligned}\mathbb{E}\left[u_i | x_i, x_{i-1}, ..., u_{t-1}, u_{t-2}, ...\right] &= 0,\\ \mathbb{E}\left[u_i^2 | x_i, x_{i-1}, ..., u_{t-1}, u_{t-2}, ...\right] &= \sigma^2.\end{aligned}$$

$$\begin{aligned}\mathbf{S} &= \sum_{j=-\infty}^{\infty}\mathbb{E}\left[(x_i u_i)(x_{i-j}u_{i-j})\right] = \sum_{j=-\infty}^{\infty}\mathbb{E}\left[u_i u_{i-j} x_i x_{i-j}\right]\\ &= ... + \mathbb{E}\left[u_i u_{i-1} x_i x_{i-1}\right] + \mathbb{E}\left[u_i^2 x_i^2\right] + \mathbb{E}\left[u_i u_{i-1} x_i x_{i+1}\right] + ...\\ &= ... + \mathbb{E}\left[\mathbb{E}\left[u_i | u_{i-1}, x_i, x_{i-1}\right] u_{i-1} x_i x_{i-1}\right] + \mathbb{E}\left[\mathbb{E}\left[u_i^2 | x_i\right] x_i^2\right] + \mathbb{E}\left[\mathbb{E}\left[u_i | u_{i+1}, x_i, x_{i+1}\right] u_{i+1} x_i x_{i+1}\right] + ...\end{aligned}$$

Remember the moment condition $\mathbb{E}[x_i u_i] = 0$. Under the assumption of independent errors (i.e., no serial correlation in u_i), which implies that all terms $j \neq 0$ are equal to zero. Thus,

$$\begin{aligned}\mathbf{S} &= \sum_{j=-\infty}^{\infty}\mathbb{E}\left[(x_i u_i)(x_{i-j}u_{i-j})\right]\\ &= \mathbb{E}\left[u_i^2 x_i^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[u_i^2 | x_i\right] x_i^2\right]\\ &= \sigma^2 \mathbb{E}\left[x_i^2\right].\end{aligned}$$

Recall that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1}\right),$$

the GMM variance-covariance matrix, $(\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1}$ is

$$\begin{aligned}\mathbf{D}^\top = \frac{\partial g}{\partial \boldsymbol{\beta}^\top} &= -\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top \mathbf{X}_i\right) = -\mathbb{E}\left[\mathbf{X}_i^\top \mathbf{X}_i\right],\\ \mathbf{S} &= \sigma^2 \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{X}_i\right],\\ \implies (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^\top)^{-1} &= \sigma^2 \mathbb{E}\left[\mathbf{X}_i^\top \mathbf{X}_i\right]^{-1}.\end{aligned}$$

Putting everything together:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathbf{D}\mathbf{S}^{-1}\mathbf{D}^{\top})^{-1}\right) = \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbb{E}\left[\mathbf{X}_i^{\top}\mathbf{X}_i\right]^{-1}\right),$$

which is equal to the standard OLS asymptotic variance-covariance matrix. Note that in the finance literature, the assumption of homoskedasticity if often violated. GMM allows for a simple correction for heteroskedasticity. The White variance-covariance matrix:

$$\begin{aligned}\mathbf{S} &= \sum_{j=-\infty}^{\infty} \mathbb{E}\left[(\mathbf{X}_i u_i)^{\top}(\mathbf{X}_{i-j} u_{i-j})\right] \\ &= \mathbb{E}\left[u_i^2 \mathbf{X}_i^{\top}\mathbf{X}_i\right],\end{aligned}$$

so that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbb{E}\left[\mathbf{X}_i^{\top}\mathbf{X}_i\right]^{-1} \mathbb{E}\left[u_i^2 \mathbf{X}_i^{\top}\mathbf{X}_i\right] \mathbb{E}\left[\mathbf{X}_i^{\top}\mathbf{X}_i\right]^{-1}\right).$$

3.5.2 Correcting standard errors for serial correlation

So far we have assumed that errors are serially uncorrelated. GMM gives us an easy way to correct the standard errors if the errors are serially correlated. The spectral density matrix is:

$$\mathbf{S} = \sum_{j=-\infty}^{\infty} \mathbb{E}\left[u_i u_{i-j} \mathbf{X}_i^{\top}\mathbf{X}_{i-j}\right].$$

If errors are serially uncorrelated, then all the $j \neq 0$ terms are zeros; if errors are serially correlated, all we need to do is to adjust $\mathbf{S}$. The most popular estimator is from Newey-West (1987):

$$\hat{\mathbf{S}}_{NW} = \sum_{j=-q}^{q} \left(\frac{q - |j|}{k}\right) \mathbb{E}\left[u_i u_{i-j} \mathbf{X}_i^{\top}\mathbf{X}_{i-j}\right].$$

$\hat{\mathbf{S}}_{NW}$ is an example of a heteroskedasticity and autocorrelation consistent (HAC) standard error. One of the powers of GMM is that it allows for an easy way to compute adjustment to standard errors – it allows us to figure out in what way errors deviate from IID assumption and compute corresponding $\mathbf{S}$.

3.5.3 Example: Estimating an MA(1) process by GMM

Consider an MA(1) process,

$$y_t = \epsilon_t + \theta \epsilon_{t-1},$$

which cannot be estimated by OLS. How about GMM? What are the moment conditions?

We started out with the model:

$$\mathbb{E}[\mathbf{h}(\theta, \mathbf{Z}_t)] = \mathbf{0},$$

and the value of minimised objective function,

$$\hat{Q} = \mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)^\top \hat{\mathbf{S}}^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t),$$

gives us an idea whether the model is "true" or not. If the model is "true", $\hat{Q}$ should be close to 0. Thus, we test whether $\hat{Q} = 0$ or not. If we reject the null, then we reject the model. We then get Hansen's J-test:

$$J = n\hat{Q} \xrightarrow{d} \chi^2(r-k),$$

which is widely used in finance and economics. It's often used as a criterion to evaluate models.

3.5.4 Example: Consumption CAPM

Suppose we have

$$\begin{aligned}
\mathbb{E}\left[\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t)\right] &= \mathbf{0}, \\
\mathbf{h}(\boldsymbol{\theta}, \mathbf{Z}_t) &= (h_1(\boldsymbol{\theta}, \mathbf{Z}_t), ..., h_J(\boldsymbol{\theta}, \mathbf{Z}_t))^\top, \\
h_j(\boldsymbol{\theta}, \mathbf{Z}_t) &= \beta \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} \frac{X_{j,t+1}}{P_{j,t}} - 1, \\
\boldsymbol{\theta} &= (\beta, \gamma)^\top.
\end{aligned}$$

The GMM estimator picks $\boldsymbol{\theta} = (\beta, \gamma)^\top$ to make

$$\hat{Q} = \mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t)^\top \hat{\mathbf{S}}^{-1} \mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t),$$

as small as possible. The vector $\mathbf{g}(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t) = (g_1(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t), ..., g_J(\hat{\boldsymbol{\theta}}, \mathbf{Z}_t))^\top$ tells us how much each moment condition deviates from 0.

$$J = T\hat{Q} \xrightarrow{d} \chi^2(r-a)$$

tells us whether we can reject the null that all moment conditions are equal to zero.

4 Simulated Method of Moments

The simulated method of moments (SMM) was first introduced into mainstream econometric discourse by McFadden (1989). As the name suggests, it is a simulation-based GMM.

Let $\{x(\omega_i, \boldsymbol{\theta})\}_{i=1}^n$ be a sequence of observed data, and let $\{x(\omega_i^s, \boldsymbol{\theta})\}_{i=1}^n$ be a sequence of simulated data for $s = 1, 2, ..., S$ and for a given parameter value $\boldsymbol{\theta}$. The simulations are done by fixing $\boldsymbol{\theta}$ and by

using the nS draws of the shocks ω_i^s. We simply write $x_i^s(\boldsymbol{\theta}) = x(\omega_i^s, \boldsymbol{\theta})$ to save on notation. Denote by $\mathbf{m}(x_i)$ a r-dimensional vector of functions of the observed data (e.g., $\mathbf{m}(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}$ if one wants to match the mean and variance of the process). The estimator for the SMM is defined as

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \left[\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{m}(x_i) - \frac{1}{S} \sum_{s=1}^{S} \mathbf{m}(x_i^s(\boldsymbol{\theta})) \right) \right]^{\top} \mathbf{W}_n \left[\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{m}(x_i) - \frac{1}{S} \sum_{s=1}^{S} \mathbf{m}(x_i^s(\boldsymbol{\theta})) \right) \right] \right\},$$

where $\mathbf{W}_n$ is the weighting matrix. If we define

$$\mathbf{h}(x_i, \boldsymbol{\theta}) = \mathbf{m}(x_i) - \frac{1}{S} \sum_{s=1}^{S} \mathbf{m}(x_i^s(\boldsymbol{\theta})), \tag{6}$$

then SMM is a special case of GMM. We can then apply the results developed for GMM.

The idea is that the true moment $\mathbb{E}[\mathbf{m}(x_i, \boldsymbol{\theta})]$ as a function of $\boldsymbol{\theta}$ is unknown. We then replace this moment with the simulated moment $\frac{1}{S} \sum_{s=1}^{S} m(x_i^s(\boldsymbol{\theta}))$.

4.1 Asymptotic properties of the SMM

Using results for GMM discussed earlier, we can establish the following properties as $n \to \infty$ for a fixed S.

The SMM estimator is consistent:

$$\hat{\boldsymbol{\theta}}_{nS} \xrightarrow{p} \boldsymbol{\theta},$$

and it is asymptotically distributed as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{nS} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_S),$$

where

$$\begin{aligned}
\boldsymbol{\Omega}_S &= \left(1 + \frac{1}{S}\right) \left[\mathbf{B}^{\top} \mathbf{W} \mathbf{B}\right]^{-1} \mathbf{B}^{\top} \mathbf{W} \boldsymbol{\Sigma} \mathbf{W} \mathbf{B} \left[\mathbf{B}^{\top} \mathbf{W} \mathbf{B}\right]^{-1}, \\
\mathbf{B} &\equiv \mathbb{E}\left[\frac{\partial \mathbf{m}(x_i^s(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \right], \quad \forall s, t, \\
\boldsymbol{\Sigma} &\equiv \operatorname*{plim}_{n \to \infty} \operatorname{Var}\left(\frac{1}{\sqrt{n}} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{m}(x_i) - \mathbb{E}\left[\mathbf{m}(x_i^s(\boldsymbol{\theta})) \right] \right) \right] \right).
\end{aligned}$$

The optimal weighting matrix is given by

$$\mathbf{W} = \left[\left(1 + \frac{1}{S}\right) \boldsymbol{\Sigma} \right]^{-1}.$$

Under this matrix, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{nS} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(1 + \frac{1}{S}\right)\left[\mathbf{B}^\top \boldsymbol{\Sigma}^{-1} \mathbf{B}\right]^{-1}\right).$$

To see how $\boldsymbol{\Omega}_S$ depends on the simulation runs, S, begin with the identity of the moment conditions and multiply with $\sqrt{n}$:

$$\begin{aligned}
\mathbf{g}(\boldsymbol{\theta}, x_i) &\equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{\theta}, x_i) \\
\sqrt{n}\mathbf{g}(\boldsymbol{\theta}, x_i) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{h}(\boldsymbol{\theta}, x_i) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{m}(x_i) - \frac{1}{S}\sum_{s=1}^{S}\mathbf{m}(x_i^s(\boldsymbol{\theta}))\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{m}(x_i) - \frac{1}{S}\sum_{s=1}^{S}\mathbf{m}(x_i^s(\boldsymbol{\theta})) + \mathbb{E}\left[\mathbf{m}(x_i)\right] - \mathbb{E}\left[\mathbf{m}(x_i)\right]\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{m}(x_i) - \mathbb{E}\left[\mathbf{m}(x_i)\right]\right) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{S}\sum_{s=1}^{S}\mathbf{m}(x_i^s(\boldsymbol{\theta})) - \mathbb{E}\left[\mathbf{m}(x_i)\right]\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbf{m}(x_i) - \mathbb{E}\left[\mathbf{m}(x_i)\right]\right) - \frac{1}{\sqrt{n}}\frac{1}{S}\sum_{i=1}^{n}\sum_{s=1}^{S}\left(\mathbf{m}(x_i^s(\boldsymbol{\theta})) - \mathbf{m}(x_i)\right),
\end{aligned}$$

where $\mathbb{E}[\mathbf{m}(x_i)]$ denotes the expectation of $\mathbf{m}(x_i)$ under the stationary distribution, and we assume a LLN with the simulation draws. Since the last two terms are independent, we can apply a CLT to derive the asymptotic distribution given above.

Empirically, we can use any consistent estimate for $\mathbf{B}$ and $\boldsymbol{\Sigma}$. In particular, we can use a HAC estimate, $\hat{\boldsymbol{\Sigma}}_{nS}$, for $\boldsymbol{\Sigma}$. We can use a J-test to test overidentifying restrictions:

$$J = n\mathbf{g}(\hat{\boldsymbol{\theta}}_{nS}, x_i)^\top \hat{\boldsymbol{\Sigma}}_{nS} \mathbf{g}(\hat{\boldsymbol{\theta}}_{nS}, x_i) \xrightarrow{d} \chi^2(r-k).$$

4.1.1 Example: McFadden's multinomial logit model

Let's consider the GMM for discrete choice models – specifically, let's look at the multinomial logit model by McFadden (1987). For discrete choice models, the GMM estimator is defined as the parameter value that solves the equations:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i(\mathbf{Z}_i, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[d_{ij} - P_{ij}(\boldsymbol{\theta})\right]\mathbf{Z}_{ij}^\top = \mathbf{0}.$$

The form can be seen as analogous to the form that the MM estimator takes for OLS:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^\top(y_i-\mathbf{X}_i\boldsymbol{\beta})=\mathbf{0},$$

or for the standard IV estimator:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i^\top(y_i-\mathbf{X}_i\boldsymbol{\beta})=\mathbf{0}.$$

For the discrete choice setting, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[d_{ij}-P_{ij}(\boldsymbol{\theta})\right]\mathbf{Z}_{ij}^\top &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i(\mathbf{Z}_i,\boldsymbol{\theta})\\ &=\mathbf{0}.\end{aligned}$$

As the conditions are not linear in $\boldsymbol{\theta}$, one uses numerical optimisation techniques to find the values of $\boldsymbol{\theta}$ that minimises the GMM criterion function:

$$Q(\boldsymbol{\theta})=\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i(\mathbf{Z}_i,\boldsymbol{\theta})\right]^\top\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i(\mathbf{Z}_i,\boldsymbol{\theta})\right].$$

There is a nice connection between the MM and the maximum likelihood approaches. Let the instruments be the gradient of the log probabilities:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[d_{ij}-P_{ij}(\boldsymbol{\theta})\right]\mathbf{Z}_{ij}^\top &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[d_{ij}-P_{ij}(\boldsymbol{\theta})\right]\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\\ &=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}d_{ij}\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}P_{ij}(\boldsymbol{\theta})\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\\ &=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}d_{ij}\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}P_{ij}(\boldsymbol{\theta})\frac{\partial P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{1}{P_{ij}(\boldsymbol{\theta})}\\ &=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}d_{ij}\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}-\frac{1}{n}\sum_{i=1}^{n}\underbrace{\sum_{j=0}^{J}\frac{\partial P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}}_{=\mathbf{0}}.\\ &=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}d_{ij}\frac{\partial\log P_{ij}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.\end{aligned}$$

The need for simulation arises when there is no closed form expression for $P_{ij}(\boldsymbol{\theta})$. In the case of simulated method of moments (SMM), we replace the exact choice probabilities, $P_{ij}(\boldsymbol{\theta})$, with simulated

probabilities, $\tilde{P}_{ij}(\boldsymbol{\theta})$. Note here that an important feature of the estimator is that $\tilde{P}_{ij}(\boldsymbol{\theta})$ enters the expression linearly. Thus, if $\tilde{P}_{ij}(\boldsymbol{\theta})$ is unbiased for $P_{ij}(\boldsymbol{\theta})$, then $\left[d_{ij} - \tilde{P}_{ij}(\boldsymbol{\theta})\right] \mathbf{Z}_{ij}^{\top}$ is unbiased for $\left[d_{ij} - P_{ij}(\boldsymbol{\theta})\right] \mathbf{Z}_{ij}^{\top}$. By not taking a non-linear transformation of the simulated probabilities, we can avoid simulation bias.

The cost of SMM is, however, a loss of efficiency. SMM is less efficient than even simulated maximum likelihood (SML), unless the ideal instruments are used. However, these ideal instruments are function of $\log \tilde{P}_{ij}(\boldsymbol{\theta})$ and thus introduce simulation bias. SMM is thus usually applied with non-ideal weights implying a loss of efficiency.

We next define the asymptotic distribution of the SMM estimator. Assume fixed instruments, so that the SMM estimator is defined by

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J}\left[d_{ij} - \tilde{P}_{ij}(\hat{\boldsymbol{\theta}})\right]\mathbf{Z}_{ij}^{\top} &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{h}}_i(\hat{\boldsymbol{\theta}}) \\ &= \tilde{\mathbf{h}}(\hat{\boldsymbol{\theta}}) \\ &= \mathbf{0}.\end{aligned}$$

Note that we can express the value of the estimating equations at the true value as

$$\begin{aligned}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0) &= \tilde{\mathbf{h}}(\boldsymbol{\theta}_0) + \left[\mathbf{h}(\boldsymbol{\theta}_0) - \mathbf{h}(\boldsymbol{\theta}_0)\right] + \left[\mathbb{E}_r\tilde{\mathbf{h}}(\boldsymbol{\theta}_0) - \mathbb{E}_r\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right] \\ &= \mathbf{h}(\boldsymbol{\theta}_0) + \left[\mathbb{E}_r\tilde{\mathbf{h}}(\boldsymbol{\theta}_0) - \mathbf{h}(\boldsymbol{\theta}_0)\right] + \left[\tilde{\mathbf{h}}(\boldsymbol{\theta}_0) - \mathbb{E}_r\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right] \\ &= \mathbf{h}(\boldsymbol{\theta}_0) + \text{Bias} + \text{Noise}.\end{aligned}$$

Rearranging the first-order Taylor expansions of $\tilde{\mathbf{h}}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ gives us:

$$\begin{aligned}\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= -\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0) \\ &= -\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\left(\mathbf{h}(\boldsymbol{\theta}_0) + \text{Bias} + \text{Noise}\right) \\ &= -\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\left(\mathbf{h}(\boldsymbol{\theta}_0) + \text{Noise}\right).\end{aligned}$$

If we have

$$\begin{aligned}\sqrt{n}\mathbf{h}(\boldsymbol{\theta}_0) &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{A}), \\ \sqrt{n}\text{Noise} &\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, R^{-1}\mathbf{S}\right),\end{aligned}$$

where $\mathbf{A}$ is the variance-covariance matrix of the non-simulated counterpart. We thus have:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\left[\mathbf{A} + R^{-1}\mathbf{S}\right]\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\right),$$

and the asymptotic distribution of the SMM estimator is then

$$\hat{\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{n}\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\left[\mathbf{A} + R^{-1}\mathbf{S}\right]\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\right).$$

The key takeaways are: the SMM variance is greater than its non-simulated counterpart. In this example, the difference is

$$\frac{1}{nR}\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1}\mathbf{S}\left(D_{\boldsymbol{\theta}}\tilde{\mathbf{h}}(\boldsymbol{\theta}_0)\right)^{-1},$$

and if R rises with n at any rate then this extra variance disappears. Finally, as non-ideal instruments are used, the estimator is less efficient than maximum likelihood.

5 The Kalman Filter

Sometimes in macroeconomics, we come across variables that play important roles in theoretical models but which we cannot observe. Examples include the concept of potential output. For example, in many Keynesian models, inflationary pressures are determined by how far actual output is from this time-varying potential output series.

In reality, we do not observe potential output so is this concept even worth bothering with? Well, just because a variable isn't observable, that doesn't mean we can't make a guess as to how it is behaving. For example, if our data moves in a way that would be consistent with a large increase in potential output (perhaps GDP rises a lot but there are no signs of inflationary pressures) then perhaps we should assume that it has indeed indeed increased.

In these notes, we will discuss methods for dealing with unobserved (or latent) variables in time series, building towards a method known as the Kalman filter. We will see the Kalman filter again when we discuss estimation of DSGE models.

Going back to our earlier point, suppose we see a big increase in output in the latest quarterly data that is not accompanied by a burst of inflation. Does this mean we should assume there has been a big change in potential output? Probably not. Potential output probably doesn't move around a lot from quarter to quarter and it is likely that there is a lot of fairly random **noise** in the quarterly fluctuations in inflation. But there is also probably a useful **signal** in the data as well.

So we are dealing with a type of signal extraction problem: What's the best way to extract a useful signal from information that also contains useless noise?

Before diving into the Kalman filter, we need some definitions and preamble, as some of the notation

used in Ljungqvist and Sargent (2018) is a bit strange.

5.1 Conditional expectations

Suppose we are interested in getting an estimate of the value of a variable, X. But we don't observe X but instead we observe a variable Z that we know to be correlated with X. Specifically, let's assume that X and Z are jointly normally distributed so that

$$\begin{bmatrix} X \\ Z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XZ} \\ \sigma_{XZ} & \sigma_Z^2 \end{bmatrix} \right).$$

In this case, the expected value of X conditional on observing Z is

$$\mathbb{E}[X|Z] = \mu_X + \frac{\sigma_{XZ}}{\sigma_Z^2}(Z - \mu_Z).$$

Alternatively, if ρ is the correlation between X and Z, $\rho = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z}$ then we can write

$$\mathbb{E}[X|Z] = \mu_X + \rho\frac{\sigma_X}{\sigma_Z}(Z - \mu_Z).$$

The amount of weight you put on the information in Z when formulating an expectation for X depends on how correlated Z is with X and on their relative standard deviation. If Z has a high standard deviation (so it's a poor signal) then you don't place much weight on it.

In the multivariate case (which we will cover in more depth next), where $\mathbf{X}$ is an n-vector and $\mathbf{Z}$ is an m-vector, there is a straightforward generalisation of the formula just presented. Denote the covariance matrix of the variables in $\mathbf{X}$ as $\mathbf{\Sigma_{XX}}$, the covariance matrix of the variables in $\mathbf{Z}$ as $\mathbf{\Sigma_{ZZ}}$, and the matrix of covariances between the entries in $\mathbf{X}$ and $\mathbf{Z}$ as $\mathbf{\Sigma_{XZ}}$.

If all the variables are jointly normally distributed, then this can be written as

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Z} \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma_{XZ}} \\ \mathbf{\Sigma_{XZ}^\top} & \mathbf{\Sigma_{ZZ}} \end{bmatrix} \right).$$

In this case, the expected value of $\mathbf{X}$ conditional on observing $\mathbf{Z}$ is

$$\mathbb{E}[\mathbf{X}|\mathbf{Z}] = \boldsymbol{\mu}_X + \mathbf{\Sigma_{XZ}}\mathbf{\Sigma_{ZZ}^{-1}}(\mathbf{Z} - \boldsymbol{\mu}_Z).$$

This formula will play an important role in our explanation of the Kalman filter.[8]

[8] In general, these conditional expectations of jointly normally distributed random variables are very hand to know in macroeconometrics, especially when looking at Bayesian estimation.

5.2 Stochastic linear difference equations

Let $\mathbf{X}_t \in \mathbb{R}^n$ denote the time t state, with a Gaussian initial distribution, $\pi_0(\mathbf{X}_0)$, with mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\Sigma}_0$, and that the transition density, $\pi(\mathbf{X}'|\mathbf{X})$, is Gaussian mean $\mathbf{AX}$ and variance $\mathbf{CC}^\top$. [9]

This specification pins down the joint distribution of the stochastic process $\{\mathbf{X}_t\}_{t=0}^{\infty}$ via

$$\pi(\mathbf{X}^t) = \pi(\mathbf{X}_t|\mathbf{X}_{t-1}) \cdots \pi(\mathbf{X}_1|\mathbf{X}_0)\pi_o(\mathbf{X}_0).$$

The joint distribution determines all moments of the process.

This specification can be represented in terms of the first-order stochastic linear difference equation

$$\mathbf{X}_{t+1} = \mathbf{AX}_t + \mathbf{CW}_{t+1}, \tag{7}$$

for $t = 0, 1, ...,$ where $\mathbf{X}_t$ is an $n \times 1$ state vector, $\mathbf{X}_0$ is a random initial condition drawn from a probability distribution with mean $\mathbb{E}[\mathbf{X}_0] = \boldsymbol{\mu}_0$ and covariance matrix $\mathbb{E}\left[(\mathbf{X}_0 - \boldsymbol{\mu}_0)(\mathbf{X}_0 - \boldsymbol{\mu}_0)^\top\right] = \boldsymbol{\Sigma}_0$, $\mathbf{A}$ is an $n \times n$ matrix, $\mathbf{C}$ is an $n \times m$ matrix, and $\mathbf{W}_{t+1}$ is an $m \times 1$ vector satisfying the following:

Assumption (A1)**.** $\mathbf{W}_{t+1}$ *is an IID process satisfying* $\mathbf{W}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

We can weaken this Gaussian assumption to focus only on first and second moments of the $\mathbf{X}$ process. It is sufficient to make the weaker assumption:

Assumption (A2)**.** $\mathbf{W}_{t+1}$ *is an* $m \times 1$ *random vector satisfying:*

$$\mathbb{E}\left[\mathbf{W}_{t+1}|\Omega_t\right] = \mathbf{0}, \tag{8}$$

$$\mathbb{E}\left[\mathbf{W}_{t+1}\mathbf{W}_{t+1}^\top|\Omega_t\right] = \mathbf{I}, \tag{9}$$

where $\Omega_t = [\mathbf{W}_t, \mathbf{W}_{t-1}, \ldots, \mathbf{W}_1, \mathbf{X}_0]$ *is the information set at* t*, and* $\mathbb{E}\left[\cdot|\Omega_t\right]$ *denotes the conditional expectation. We impose no distributional assumptions beyond this assumption. A sequence* $\{\mathbf{W}_{t+1}\}$ *satisfying (8) is said to be a Martingale difference sequence (MDS) adapted to* Ω_t*.*[10]

An even weaker assumption is:

Assumption (A3)**.** $\mathbf{W}_{t+1}$ *is a process satisfying*

$$\mathbb{E}[\mathbf{W}_{t+1}] = \mathbf{0},$$

[9] An $n \times 1$ vector, $\mathbf{z}$, that is multivariate normal has the density function

$$\phi(\mathbf{z}) = (2\pi)^{-\frac{1}{2}n}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{z} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{z} - \boldsymbol{\mu}\right)\right\},$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{z}]$ and $\boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^\top\right]$.

[10] Note that (8) by itself allows $\mathbf{W}_{t+1}$ conditional on Ω_t to be heteroskedastic.

for all t, and

$$\mathbb{E}[\mathbf{W}_t \mathbf{W}_{t-j}^\top] = \begin{cases} \mathbf{I}, & \text{if } j = 0, \\ \mathbf{O} & \text{if } j \neq 0. \end{cases}$$

A process satisfying this assumption is said to be a vector of "white noise".

Assumption A1 or A2 implies A3 but not vice versa. Assumption A1 implies A2 but not vice versa. Assumption A3 is sufficient to justify the formulas that we report below for second moments. We shall often append an observation equation $\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t$ to Equation (7) and deal with the augmented system:

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1}, \tag{10}$$

$$\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t. \tag{11}$$

Here $\mathbf{Y}_t$ is a vector variables observed at t, which may include some linear combinations of $\mathbf{X}_t$. The system made up of (10) and (11) is often called a **linear state-space system**.

5.2.1 Example: Scalar second-order autoregression

Assume that z_t and w_t are scalar processes and that

$$z_{t+1} = \alpha + \rho_1 z_t + \rho_2 z_{t-1} + w_{t+1}.$$

Represent this relationship as the system

$$\begin{bmatrix} z_{t+1} \\ z_t \\ 1 \end{bmatrix} = \begin{bmatrix} \rho_1 & \rho_2 & \alpha \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_t \\ z_{t-1} \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} w_t,$$

$$z_t = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ z_{t-1} \\ 1 \end{bmatrix},$$

which has the form of (10) and (11).

5.2.2 Example: First-order scalar mixed moving average and autoregression

Let

$$z_{t+1} = \rho z_t + w_{t+1} + \gamma w_t.$$

Express this relationship as

$$\begin{bmatrix} z_{t+1} \\ w_{t+1} \end{bmatrix} = \begin{bmatrix} \rho & \gamma \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ w_t \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} w_{t+1},$$
$$z_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} z_t \\ w_t \end{bmatrix}.$$

5.2.3 Example: Vector autoregression

Let $\mathbf{Z}_t$ be an $n \times 1$ vector random variables. We define the VAR(4) system by a stochastic difference equation

$$\mathbf{Z}_{t+1} = \sum_{j=1}^{4} \mathbf{A}_j \mathbf{Z}_{t+1-j} + \mathbf{C}_y \mathbf{W}_{t+1}, \tag{12}$$

where $\mathbf{W}_{t+1}$ is a MDS satisfying (8)-(9) with

$$\mathbf{X}_0' = \begin{bmatrix} \mathbf{Z}_0 & \mathbf{Z}_{-1} & \mathbf{Z}_{-2} & \mathbf{Z}_{-3} \end{bmatrix},$$

and $\mathbf{A}_j$ is an $n \times n$ matrix for each j. We can map this VAR into Equation (7) as follows:

$$\begin{bmatrix} \mathbf{Z}_{t+1} \\ \mathbf{Z}_t \\ \mathbf{Z}_{t-1} \\ \mathbf{Z}_{t-2} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{A}_3 & \mathbf{A}_4 \\ \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \mathbf{Z}_t \\ \mathbf{Z}_{t-1} \\ \mathbf{Z}_{t-2} \\ \mathbf{Z}_{t-3} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_y \\ \mathbf{O} \\ \mathbf{O} \\ \mathbf{O} \end{bmatrix} \mathbf{W}_{t+1}, \tag{13}$$

where $\mathbf{O}$ denotes a null matrix with the right dimensions for conformity. Define $\mathbf{A}$ as the state transition matrix, and assume that $\mathbf{A}$ has all of its eigenvalues bounded in modulus below unity (within unit circle). Then (12) can be initialised so that $\mathbf{Z}_t$ is covariance stationary.

5.3 First and second moments

We can use Equation (7) to deduce the first and second moments of the sequence of random vectors $\{\mathbf{X}_t\}_{t=0}^{\infty}$. A sequence of random vectors is called a **stochastic process**.

Definition (LS 2.4.1)**.** A stochastic process, $\{\mathbf{X}_t\}$, is said to be covariance stationary if it satisfies the following two properties:

1. The mean is independent of time, $\mathbb{E}[\mathbf{X}_t] = \mathbb{E}[\mathbf{X}_0],\ \forall t$.
2. The sequence of autocovariance matrices, $\mathbb{E}\left[(\mathbf{X}_{t+j} - \mathbb{E}[\mathbf{X}_{t+j}])(\mathbf{X}_t - \mathbb{E}[\mathbf{X}_t])^\top\right]$ depends on the separation between dates $j = 0, \pm 1, \pm 2, ...$, but not on t.

We use the following definition too:

Definition (LS 2.4.2)**.** A square real valued matrix $\mathbf{A}$ is said to be stable if all of its eigenvalues modulus are strictly less than unity.[11]

We shall often find it useful to assume that (10)-(11) takes the special form

$$\begin{bmatrix} \mathbf{X}_{1,t+1} \\ \mathbf{X}_{2,t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\iota}\boldsymbol{\iota}^\top & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{A}} \end{bmatrix}}_{\mathbf{A}} \begin{bmatrix} \mathbf{X}_{1,t} \\ \mathbf{X}_{2,t} \end{bmatrix} + \begin{bmatrix} \mathbf{O} \\ \tilde{\mathbf{C}} \end{bmatrix} \mathbf{W}_{t+1}, \tag{14}$$

where $\boldsymbol{\iota}$ is the $n \times 1$ unit vector, and $\tilde{\mathbf{A}}$ is a stable matrix. That $\tilde{\mathbf{A}}$ is a stable matrix implies that the only solution of $(\tilde{\mathbf{A}} - \mathbf{I})\boldsymbol{\mu}_2 = \mathbf{0}$ is $\boldsymbol{\mu}_2 = \mathbf{0}$ (i.e., 1 is not a stable eigenvalue of $\tilde{\mathbf{A}}$). It follows that the matrix $\mathbf{A}$ on the RHS of (14) has one eigenvector associated with a single unit eigenvalue:

$$(\mathbf{A} - \mathbf{I}) \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \mathbf{0},$$

which implies $\boldsymbol{\mu}_1$ is an arbitrary vector and $\boldsymbol{\mu}_2 = \mathbf{0}$. The first equation of (14) implies $\mathbf{X}_{1,t+1} = \mathbf{X}_{1,0},\ \forall t \geq 0$. Picking the initial condition $\mathbf{X}_{1,0}$ pins down a particular eigenvector, $\begin{bmatrix} \mathbf{X}_{1,0} \\ \mathbf{0} \end{bmatrix}$, of $\mathbf{A}$. As we shall soon see, this eigenvector is our candidate for the unconditional mean of $\mathbf{X}$ that makes the process covariance stationary.

We will make an assumption that guarantees that there exists an initial condition,

$$(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \left(\mathbb{E}[\mathbf{X}_0], \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}_0]) (\mathbf{X} - \mathbb{E}[\mathbf{X}_0])^\top \right] \right), \tag{15}$$

that makes the $\mathbf{X}_t$ process covariance stationary. Either of the following conditions works:

Condition (C1)**.** All of the eigenvalues of $\mathbf{A}$ in (10)-(11) are strictly less than 1 in modulus.

Condition (C2)**.** The state-space representation takes the special form (14) and all of the eigenvalues of $\tilde{\mathbf{A}}$ are strictly less than 1 in modulus.

To discover the first and second moments of the $\mathbf{X}_t$ process, we regard the initial condition $\mathbf{X}_0$ as being drawn from a distribution with mean $\boldsymbol{\mu}_0 = \mathbb{E}[\mathbf{X}_0]$ and covariance $\boldsymbol{\Sigma}_0 = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}_0]) (\mathbf{X} - \mathbb{E}[\mathbf{X}_0])^\top \right]$. We shall deduce starting values for the mean and covariance that make the process covariance stationary, though our formulas are also useful for describing what happens when we start from other initial conditions that generate transient behaviour that stops the process from being covariance stationary.

[11] This is related to the Blanchard-Kahn conditions for determinacy in DSGE models.

Taking mathematical expectations on both sides of Equation (7) gives

$$\begin{aligned}\mathbb{E}\left[\mathbf{X}_{t+1}\right] &= \mathbb{E}\left[\mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1}\right] \\ \boldsymbol{\mu}_{t+1} &= \mathbf{A}\boldsymbol{\mu}_t, \end{aligned} \tag{16}$$

where $\boldsymbol{\mu}_t = \mathbb{E}[\mathbf{X}_t]$. We will assume that all of the eigenvalues of $\mathbf{A}$ are strictly less than unity in modulus, except possible for one that is affiliated with the constant terms in the various equations. Then, $\mathbf{X}_t$ possesses a stationary mean defined to satisfy $\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t$, which from Equation (16) evidently satisfies

$$(\mathbf{I} - \mathbf{A})\boldsymbol{\mu} = \mathbf{0}, \tag{17}$$

which characterises the mean, $\boldsymbol{\mu}$, as an eigenvector associated with the single unit eigenvalue of $\mathbf{A}$. The condition that the remaining eigenvalues of $\mathbf{A}$ are less than unity in modulus implies that starting from any $\boldsymbol{\mu}_0, \boldsymbol{\mu}_t \to \boldsymbol{\mu}$.[12]

Notice that

$$\mathbf{X}_{t+1} - \boldsymbol{\mu}_{t+1} = \mathbf{A}(\mathbf{X}_t - \boldsymbol{\mu}_t) + \mathbf{C}\mathbf{W}_{t+1}. \tag{18}$$

From this equation, we can compute the law of motion of the covariance matrices $\boldsymbol{\Sigma}_t \equiv \mathbb{E}\left[\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)^\top\right]$. Thus,

$$\mathbb{E}\left[\left(\mathbf{X}_{t+1} - \boldsymbol{\mu}_{t+1}\right)\left(\mathbf{X}_{t+1} - \boldsymbol{\mu}_{t+1}\right)^\top\right] = \mathbf{A}\mathbb{E}\left[\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)^\top\right]\mathbf{A}^\top + \mathbf{C}\underbrace{\mathbb{E}\left[\mathbf{W}_{t+1}\mathbf{W}_{t+1}^\top\right]}_{=\mathbf{I}}\mathbf{C}^\top,$$

or

$$\boldsymbol{\Sigma}_{t+1} = \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top.$$

A fixed point of this matrix difference equation evidently satisfies

$$\boldsymbol{\Sigma}_\infty = \mathbf{A}\boldsymbol{\Sigma}_\infty\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top. \tag{19}$$

A fixed point, $\boldsymbol{\Sigma}_\infty$, is the covariance matrix $\mathbb{E}\left[\left(\mathbf{X}_t - \boldsymbol{\mu}\right)\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)^\top\right]$ under a stationary distribution of $\mathbf{X}$. Equation (19) is a **discrete Lyapunov equation** in the $n \times n$ matrix $\boldsymbol{\Sigma}_\infty$.[13]

[12] To understand this, assume that the eigenvalues of $\mathbf{A}$ are distinct, and use the Jordan decomposition, $\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{A}$, arranged in descending order of magnitude, and $\mathbf{P}$ is a matrix composed of the corresponding eigenvectors. Then Equation (16) can be expressed as $\boldsymbol{\mu}^*_{t+1} = \boldsymbol{\Lambda}\boldsymbol{\mu}^*_t$, where $\boldsymbol{\mu}^*_t = \mathbf{P}^{-1}\boldsymbol{\mu}_t$, which implies that $\boldsymbol{\mu}^*_t = \boldsymbol{\Lambda}^t\boldsymbol{\mu}^*_0$. When all eigenvalues but the first are less than unity, $\boldsymbol{\Lambda}^t$ converges to matrix of zeroes except for the $(1,1)$ element, and $\boldsymbol{\mu}^*_t$ converges to a vector of zeroes except for the first element, which stays at $\boldsymbol{\mu}^*_{0,1}$, its initial value, which we are free to set equal to 1, to capture the constant. Then $\boldsymbol{\mu}_t = \mathbf{P}\boldsymbol{\mu}^*_t$ converges to $\mathbf{P}_1\boldsymbol{\mu}^*_{0,1} = \mathbf{P}_1$, where $\mathbf{P}_1$ is the eigenvector corresponding to the unit eigenvalue.

[13] It can be solved using the `doublej.m` file from Ljungqvist and Sargent 2018.

By virtue of (7) and (16), note that for $j \geq 0$

$$(\mathbf{X}_{t+j} - \boldsymbol{\mu}_{t+j}) = \mathbf{A}^j(\mathbf{X}_t - \boldsymbol{\mu}_t) + \mathbf{C}\mathbf{W}_{t+j} + \cdots + \mathbf{A}^{j-1}\mathbf{C}\mathbf{W}_{t+1}.$$

Postmultiplying both sides by $(\mathbf{X}_t - \boldsymbol{\mu}_t)^\top$ and taking expectations shows that the autocovariance sequence satisfies

$$\boldsymbol{\Sigma}_{t+j,t} \equiv \mathbb{E}\left[\left(\mathbf{X}_{t+j} - \boldsymbol{\mu}_{t+j}\right)\left(\mathbf{X}_t - \boldsymbol{\mu}_t\right)^\top\right] = \mathbf{A}^j\boldsymbol{\Sigma}_t. \tag{20}$$

Note that $\boldsymbol{\Sigma}_{t+j,t}$ depends on both j, the gap between dates, and t, the earlier date.

In the special case that $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_\infty$ that solves the Lyapunov equation (19), $\boldsymbol{\Sigma}_{t+j,t} = \mathbf{A}_0^j\boldsymbol{\Sigma}_\infty$ and so depends on the gap j between time periods. In this case, an autocovariance matrix sequence $\{\boldsymbol{\Sigma}_{t+j,t}\}_{j=0}^\infty$ is often called an **autocovariogram**.

Suppose that $\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t$. Then $\boldsymbol{\mu}_{y,t} = \mathbb{E}[\mathbf{Y}_t] = \mathbf{G}\boldsymbol{\mu}_t$ and

$$\mathbb{E}\left[\left(\mathbf{Y}_{t+j} - \boldsymbol{\mu}_{y,t+j}\right)\left(\mathbf{Y}_t - \boldsymbol{\mu}_{y,t}\right)^\top\right] = \mathbf{G}\Sigma_{t+j,t}\mathbf{G}^\top, \tag{21}$$

for $j = 0, 1, \ldots$. Equations (21) show that the autocovariogram for a stochastic process governed by a stochastic linear difference equation obeys the nonstochastic version of that difference equation.

5.4 Population regression

This section explains the notion of a population regression equation. Suppose that we have a state-space system as in (10)-(11),

$$\begin{aligned}\mathbf{X}_{t+1} &= \mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1},\\ \mathbf{Y}_t &= \mathbf{G}\mathbf{X}_t,\end{aligned}$$

with initial conditions that make it covariance stationary. We can use the preceding formulas to compute the second moments of any pair of random variables. These moments let us compare a linear regression. Thus, let $\mathbf{X}$ be a $1 \times p$ vector of random variables somehow selected from the stochastic process $\{\mathbf{Y}_t\}$ governed by the system (10)-(11).

For example, let $p = 2m$, where $\mathbf{Y}_t$ is an $m \times 1$ vector, and take $\mathbf{X} = \begin{bmatrix}\mathbf{Y}_t^\top & \mathbf{Y}_{t-1}^\top\end{bmatrix}$ for any $t \geq 1$. Let Y be any scalar random variable selected from the $m \times 1$ stochastic process $\{\mathbf{Y}_t\}$. For example, take $Y = \mathbf{Y}_{t+1,1}$ for the same t used to define $\mathbf{X}$, where $\mathbf{Y}_{t+1,1}$ is the first component of $\mathbf{Y}_{t+1}$.

We consider the following least-squares approximation problem: find a $p \times 1$ vector real numbers, $\boldsymbol{\beta}$, that attain

$$\arg\min_{\boldsymbol{\beta}} \mathbb{E}\left[Y - \mathbf{X}\boldsymbol{\beta}\right]^2. \tag{22}$$

Here $\mathbf{X}\boldsymbol{\beta}$ is being used to estimate Y, and we want the value of $\boldsymbol{\beta}$ that minimises the expected squared

error. The FOC is:

$$\mathbb{E}\left[\mathbf{X}^\top(Y - \mathbf{X}\boldsymbol{\beta})\right] = \mathbf{0}, \tag{23}$$

which can be rearranged as

$$\boldsymbol{\beta} = \left(\mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\right]\right)^{-1}\mathbb{E}\left[\mathbf{X}^\top Y\right]. \tag{24}$$

By using the formulas (17), (19), (20), and (21), we can compute $\mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\right]$ and $\mathbb{E}\left[\mathbf{X}^\top Y\right]$ for whatever selection of $\mathbf{X}$ and Y we choose. The condition (23) is called the least-squares normal equation – this should all be very familiar – and it states that the projection error $Y - \mathbf{X}\boldsymbol{\beta}$ is orthogonal to $\mathbf{X}$. Therefore, we can represent Y as

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{25}$$

where $\mathbb{E}[\mathbf{X}^\top\epsilon] = \mathbf{0}$. The above equation is called a population regression equation. The vector $\boldsymbol{\beta}$ is called the population least-squares regression vector. The law of large numbers for continuous-state Markov processes states conditions that guarantee that sample moments converge to population moments – i.e.,

$$\frac{1}{S}\sum_{s=1}^{S}\mathbf{X}_s^\top\mathbf{X}_s \to \mathbb{E}[\mathbf{X}^\top\mathbf{X}],$$
$$\frac{1}{S}\sum_{s=1}^{S}\mathbf{X}_s^\top Y_s \to \mathbb{E}[\mathbf{X}^\top Y].$$

Under these conditions, sample least-squares estimates converge to $\boldsymbol{\beta}$.

5.4.1 Multiple regressors

Now let $\mathbf{Y}$ be an $n \times 1$ vector of random variables and think of regression solving the least-squares problem for each of them to attain a representation

$$\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X}^\top + \boldsymbol{\epsilon}, \tag{26}$$

where $\boldsymbol{\Gamma}$ is now $n \times p$ and $\boldsymbol{\epsilon}$ is $n \times 1$ vector least squares residuals. The population regression coefficients are now given by

$$\arg\min_{\boldsymbol{\Gamma}} \mathbb{E}\left[\mathbf{Y} - \boldsymbol{\Gamma}\mathbf{X}^\top\right],$$

with the following FOC:

$$\mathbb{E}\left[(\mathbf{Y} - \boldsymbol{\Gamma}\mathbf{X}^\top)\mathbf{X}\right] = \mathbf{0}.$$

This yields

$$\boldsymbol{\Gamma} = \mathbb{E}\left[\mathbf{Y}\mathbf{X}\right]\left(\mathbb{E}\left[\mathbf{X}^\top\mathbf{X}\right]\right)^{-1}. \tag{27}$$

We will use this formula repeatedly to derive the Kalman filter.

5.5 Deriving the Kalman filter

As a fruitful application of the population regression formula (27), we derive the celebrated Kalman filter for the state space system for $t \geq 0$:

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1}, \tag{28}$$

$$\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t + \mathbf{V}_t, \tag{29}$$

where $\mathbf{X}_t$ is an $n \times 1$ (hidden) state vector and $\mathbf{Y}_t$ is an $m \times 1$ vector of signals on the hidden state; $\mathbf{W}_{t+1}$ is a $p \times 1$ vector IID distributed as $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. We assume that $\mathbf{W}_{t+1}$ and $\mathbf{V}_s$ are orthogonal (i.e., $\mathbb{E}[\mathbf{W}_{t+1}\mathbf{V}_s^\top] = \mathbf{O}$) for all $t+1$ and s greater than or equal to 0. We assume that

$$\mathbf{X}_0 \sim \mathcal{N}(\hat{\mathbf{X}}_0, \mathbf{\Sigma}_0). \tag{30}$$

We assume that we observe $\mathbf{Y}_t, ..., \mathbf{Y}_0$ but not $\mathbf{X}_t, ..., \mathbf{X}_0$ at time t. We know all first and second moments implied by the structure (28)-(30).

System (28)-(30) is an example of a hidden Markov model. The stochastic process $\{\mathbf{Y}_t\}_{t=0}^{\infty}$ is not Markov, but the hidden process $\{\mathbf{X}_t\}_{t=0}^{\infty}$ is Markov, and so is the process $\{\hat{\mathbf{X}}_t, \mathbf{\Sigma}_t\}$ that constitutes sufficient statistics for the probability distributions of $\mathbf{Y}_t$ conditional on $[\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, ..., \mathbf{Y}_0]$.

We work forward in time, start time $t = 0$ before we observe $\mathbf{Y}_0$. Specification (29) and (30) implies that the condition distribution of $\mathbf{Y}_0$ is

$$\mathbf{Y}_0 \sim \mathcal{N}\left(\mathbf{G}\hat{\mathbf{X}}_0, \mathbf{G}\mathbf{\Sigma}_0\mathbf{G}^\top + \mathbf{R}\right), \tag{31}$$

which recall comes from (15). For $t \geq 0$, let $\mathbf{Y}^t = [\mathbf{Y}_t, \mathbf{Y}_{t-1}, ..., \mathbf{Y}_0]$. We want to seek an expression for the probability distribution of $\mathbf{Y}_t$ condition on history $\mathbf{Y}^{t-1}$ that has a convenient recursive representation. The Kalman filter attains that by constructing recursive formulas for objects $(\hat{\mathbf{X}}_t, \mathbf{\Sigma}_t)$ that appear in the following generalisation of (31)

$$\mathbf{Y}_t \sim \mathcal{N}\left(\mathbf{G}\hat{\mathbf{X}}_t, \mathbf{G}\mathbf{\Sigma}_t\mathbf{G}^\top + \mathbf{R}\right). \tag{32}$$

The objects $(\hat{\mathbf{X}}_t, \mathbf{\Sigma}_t)$ characterise the population regression $\hat{\mathbf{X}}_t = \mathbb{E}[\mathbf{X}_t|\mathbf{Y}_{t-1}, ..., \mathbf{Y}_0]$ and the covariance $\mathbf{\Sigma}_t = \mathbb{E}\left[(\mathbf{X}_t - \hat{\mathbf{X}}_t)(\mathbf{X}_t - \hat{\mathbf{X}}_t)^\top\right]$.

At each date, our approach is to regress what we don't know on what we know. Let's start $t = 0$. We arrive at date 0 knowing $\hat{\mathbf{X}}_0, \mathbf{\Sigma}_0$. Then we observe $\mathbf{Y}_0$ and make some inferences. It will turn out that among the objects with which we leave time $t = 0$ will be $\hat{\mathbf{X}}_1, \mathbf{\Sigma}_1$. This gives a perspective from

which "we are in the same situation at the start of period 1 that we were at the start of period 0", an insight that activates a recursion.

We use the insight that the information in $\mathbf{Y}_0$ that is new relative to the information $(\hat{\mathbf{X}}_0, \mathbf{\Sigma}_0)$ that we knew before observing $\mathbf{Y}_0$ is $\mathbf{A}_0 \equiv \mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0$. Thus, before we observe $\mathbf{Y}_0$, we regard $\mathbf{X}_0$ as a random vector with mean $\hat{\mathbf{X}}_0$ and covariance matrix $\mathbf{\Sigma}_0$. Then we observe the random vector $\mathbf{Y}_0$ linked to $\mathbf{X}_0$ by the time 0 version of Equation (29). We form revised beliefs about the mean of $\mathbf{X}_0$ after observing $\mathbf{Y}_0$ by computing the distribution of $\mathbf{X}_0$ conditional on $\mathbf{Y}_0$. The conditional mean $\mathbb{E}[\mathbf{X}_0|\mathbf{Y}_0] = \hat{\mathbf{X}}_0 + \mathbf{L}_0(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0)$ satisfies the appropriate version of the population regression formula (27), namely,

$$\mathbf{X}_0 - \hat{\mathbf{X}}_0 = \mathbf{L}_0(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0) + \boldsymbol{\eta}, \tag{33}$$

where $\boldsymbol{\eta}$ is vector of least squares residuals whose orthogonality to $(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0)$ characterises $\mathbf{L}_0$ as population least squares regression coefficients. The least squares orthogonality conditions are

$$\mathbb{E}\left[(\mathbf{X}_0 - \hat{\mathbf{X}}_0)(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0)^\top\right] = \mathbf{L}_0\mathbb{E}\left[(\mathbf{X}_0 - \mathbf{G}\hat{\mathbf{X}}_0)(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0)^\top\right].$$

Evaluating the moment matrices and solving for $\mathbf{L}_0$ gives the formula

$$\mathbf{L}_0 = \mathbf{\Sigma}_0\mathbf{G}^\top(\mathbf{G}\mathbf{\Sigma}_0\mathbf{G}^\top + \mathbf{R})^{-1}. \tag{34}$$

Having constructed $\mathbb{E}[\mathbf{X}_0|\mathbf{Y}_0]$, we can construct $\hat{\mathbf{X}}_1 = \mathbb{E}[\mathbf{X}_1|\mathbf{Y}_0]$ as follows.[14] Equation (28) implies $\mathbb{E}[\mathbf{X}_1|\hat{\mathbf{X}}_0] = \mathbf{A}\mathbf{X}_0$ and that

$$\mathbf{X}_1 = \mathbf{A}\hat{\mathbf{X}}_0 + \mathbf{A}(\mathbf{X}_0 - \hat{\mathbf{X}}_0) + \mathbf{C}\mathbf{W}_1. \tag{35}$$

Furthermore, applying (33) shows that $\mathbb{E}[\mathbf{X}_1|\mathbf{Y}_0] = \mathbf{A}\hat{\mathbf{X}}_0 + \mathbf{A}\mathbf{L}_0(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0)$, which we express as

$$\hat{\mathbf{X}}_1 = \mathbf{A}\hat{\mathbf{X}}_0 + \mathbf{K}_0(\mathbf{Y} - \mathbf{G}\hat{\mathbf{X}}_0), \tag{36}$$

where

$$\mathbf{K}_0 = \mathbf{A}\mathbf{\Sigma}_0\mathbf{G}^\top(\mathbf{G}\mathbf{\Sigma}_0\mathbf{G}^\top + \mathbf{R})^{-1}.$$

Subtract (36) from (35) to get

$$\mathbf{X}_1 - \hat{\mathbf{X}}_1 = \mathbf{A}(\mathbf{X}_0 - \hat{\mathbf{X}}_0) + \mathbf{C}\mathbf{W}_1 - \mathbf{K}_0(\mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0). \tag{37}$$

Use this equation and $\mathbf{Y}_0 = \mathbf{G}\mathbf{X}_0 + \mathbf{V}_0$ to compute the following formula for the conditional variance,

[14] It is understood that we know $\hat{\mathbf{X}}_0$. Instead of writing $\mathbb{E}[\mathbf{X}_1|\mathbf{Y}_0, \hat{\mathbf{X}}_0]$, we choose simply to write $\mathbb{E}[\mathbf{X}_1|\mathbf{Y}_0]$, but we intend the meaning to be the same. More generally, when we write $\mathbb{E}[\mathbf{X}_t|\mathbf{Y}^{t-1}]$, it understood that the mathematical expectation is also condition on $\hat{\mathbf{X}}_0$.

$\mathbb{E}\left[(\mathbf{X}_1 - \hat{\mathbf{X}}_1)(\mathbf{X}_1 - \hat{\mathbf{X}}_1)^\top\right] = \mathbf{\Sigma}_1$:

$$\mathbf{\Sigma}_1 = (\mathbf{A} - \mathbf{K}_0\mathbf{G})\mathbf{\Sigma}_0(\mathbf{A} - \mathbf{K}_0\mathbf{G})^\top + (\mathbf{C}\mathbf{C}^\top + \mathbf{K}_0\mathbf{R}\mathbf{K}_0^\top). \tag{38}$$

Thus, we have deduced the conditional distribution, $\mathbf{X}_1|\mathbf{Y}_0 \sim \mathcal{N}(\hat{\mathbf{X}}_1, \mathbf{\Sigma}_1)$. Collecting equations, we can write

$$\mathbf{a}_0 = \mathbf{Y}_0 - \mathbf{G}\hat{\mathbf{X}}_0, \tag{39}$$
$$\mathbf{K}_0 = \mathbf{A}\mathbf{\Sigma}_0\mathbf{G}^\top(\mathbf{G}\mathbf{\Sigma}_0\mathbf{G}^\top + \mathbf{R})^{-1}, \tag{40}$$
$$\hat{\mathbf{X}}_1 = \mathbf{A}\hat{\mathbf{X}}_0 + \mathbf{K}_0\mathbf{a}_0, \tag{41}$$
$$\mathbf{\Sigma}_0 = \mathbf{C}\mathbf{C}^\top + \mathbf{K}_0\mathbf{R}\mathbf{K}_0^\top + (\mathbf{A} - \mathbf{K}_0\mathbf{G})\mathbf{\Sigma}_0(\mathbf{A} - \mathbf{K}_0\mathbf{G})^\top. \tag{42}$$

Among the outcomes of system (39)-(42) is a conditional mean, covariance pair $(\hat{\mathbf{X}}_1, \mathbf{\Sigma}_1)$. It is appropriate to view system (39)-(42) as a mapping of a mean, covariance pair $(\hat{\mathbf{X}}_0, \mathbf{\Sigma}_0)$ into a mean, and a covariance pair $(\hat{\mathbf{X}}_1, \mathbf{\Sigma}_1)$, with auxiliary intermediate outputs $(\mathbf{a}_0, \mathbf{K}_0)$. The Kalman filter iterates on this mapping to arrive at the following recursions for $t \geq 0$:

$$\mathbf{a}_t = \mathbf{Y}_t - \mathbf{G}\hat{\mathbf{X}}_t, \tag{43}$$
$$\mathbf{K}_t = \mathbf{A}\mathbf{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\mathbf{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}, \tag{44}$$
$$\hat{\mathbf{X}}_{t+1} = \mathbf{A}\hat{\mathbf{X}}_t + \mathbf{K}_t\mathbf{a}_t, \tag{45}$$
$$\mathbf{\Sigma}_t = \mathbf{C}\mathbf{C}^\top + \mathbf{K}_t\mathbf{R}\mathbf{K}_t^\top + (\mathbf{A} - \mathbf{K}_t\mathbf{G})\mathbf{\Sigma}_t(\mathbf{A} - \mathbf{K}_t\mathbf{G})^\top. \tag{46}$$

System (43)-(46) is the **Kalman filter**, and $\mathbf{K}_t$ is called the **Kalman gain**. Substituting for $\mathbf{K}_t$ from (44) allows us to rewrite (46) as

$$\mathbf{\Sigma}_{t+1} = \mathbf{A}\mathbf{\Sigma}_t\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top - \mathbf{A}\mathbf{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\mathbf{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}\mathbf{G}\mathbf{\Sigma}_t\mathbf{A}^\top. \tag{47}$$

This equation is known as a matrix **Riccati difference equation** that restricts a sequence of covariance matrices $\{\mathbf{\Sigma}_t\}_{t=0}^\infty$.

5.5.1 The Kalman smoother

The Kalman filter is what is known as a **one-sided** filter: The estimates of states at time t are based solely on information available at time t. No data after period t is used to calculate estimates of the unobserved state variables. This is a reasonable model for how someone might behave if they were learning about the state variables in real time. But researchers have access to the full history of the data set, including all observations after time t.

For this reason, some macroeconomists generally estimate time-varying models using a method

known as the Kalman smoother. This is a **two-sided** filter that uses data both before and after time t to compute expected values of the state variables at time t.[15]

5.6 Vector autoregressions and the Kalman filter

5.6.1 Conditioning on the semi-infinite past of Y

Under conditions summarised by Anderson et al. (1996), iterations on (44), (46) converge to time-invariant $\mathbf{K}, \mathbf{\Sigma}$ for any positive semidefinite initial covariance matrix $\mathbf{\Sigma}_0$. A time-invariant matrix $\mathbf{\Sigma}_t = \mathbf{\Sigma}$ that solves (46) is the covariance matrix of $\mathbf{X}_t$ around $\mathbb{E}\left[\mathbf{X}_t | \{\mathbf{Y}_{-\infty}^{t-1}\}\right]$, where $\{\mathbf{Y}_{-\infty}^{t-1}\}$ denotes the semi-infinite history of $\mathbf{Y}_s$ for all dates on or before $t-1$.[16]

5.6.2 A time-invariant VAR

Suppose that the fixed point of (46) just described exists. If we initiate (46) from this fixed point $\mathbf{\Sigma}$, then the innovations representations becomes time invariant:

$$\hat{\mathbf{X}}_{t+1} = \mathbf{A}\hat{\mathbf{X}}_t + \mathbf{K}\mathbf{a}_t, \tag{48}$$

$$\mathbf{Y}_t = \mathbf{G}\hat{\mathbf{X}}_t + \mathbf{a}_t, \tag{49}$$

where $\mathbb{E}[\mathbf{a}_t \mathbf{a}_t^\top] = \mathbf{G}\mathbf{\Sigma}\mathbf{G}^\top + \mathbf{R}$. Use (48) and (49) to express $\hat{\mathbf{X}}_{t+1} = (\mathbf{A} - \mathbf{K}\mathbf{G})\hat{\mathbf{X}}_t + \mathbf{K}\mathbf{Y}_t$. If we assume that the eigenvalues of $\mathbf{A} - \mathbf{K}\mathbf{G}$ are bounded in modulus below unity,[17] we can solve the preceding equation to get

$$\hat{\mathbf{X}}_{t+1} = \sum_{j=0}^{\infty} (\mathbf{A} - \mathbf{K}\mathbf{G})^j \mathbf{K}\mathbf{Y}_{t-j}.$$

Then solving (49) for $\mathbf{Y}_t$ gives the VAR

$$\mathbf{Y}_t = \mathbf{G}\sum_{j=0}^{\infty} (\mathbf{A} - \mathbf{K}\mathbf{G})^j \mathbf{K}\mathbf{Y}_{t-j-1} + \mathbf{a}_t, \tag{50}$$

where by construction

$$\mathbb{E}[\mathbf{a}_t \mathbf{Y}_{t-j-1}^\top] = \mathbf{O} \quad j \geq 0.$$

The orthogonality conditions identity (50) as a VAR.

[15] Although, as explained by Pfeifer (2013), one should not use a two-sided filter when preparing data for a DSGE model when using Dynare (with some exceptions, such as population stats). This is because of the backwards looking nature of the state space solution system that Dynare constructs.

[16] The Matlab program `kfilter.m` from Ljungqvist and Sargent (2018) implements the time-invariant Kalman filter, allowing for correlation between $\mathbf{W}_{t+1}$ and $\mathbf{V}_t$.

[17] Anderson et al. (1996) show assumptions that guarantee that the eigenvalues of $\mathbf{A} - \mathbf{K}\mathbf{G}$ are bounded within unit circle.

5.7 Applications of the Kalman filter

5.7.1 Muth's reverse engineering exercise

Cagan (1956) and Friedman (1957) posited that to form expectations of future values of a scalar y_t, people use the following "adaptive expectations" scheme:

$$y^*_{t+1} = K \sum_{j=0}^{\infty} (1-K)^j y_{t-j},$$
$$\Leftrightarrow y^*_{t+1} = (1-K) y^*_t + K y_t,$$

where y^*_{t+1} is the public's expectation. Friedman used this scheme to describe people's forecasts of future income. Cagan used it to model their forecasts of inflation during hyperinflations. Cagan and Friedman did not asset that the scheme is an optimal one, and so did not fully defend it. Muth (1960) wanted to understand the circumstances under which this forecasting scheme would be optimal. Therefore, he sought a stochastic process for y_t such that the aforementioned equations would be optimal. In effect, he posed and solved an "inverse optimal prediction" problem of the form "you give me the forecasting scheme; I'll find the stochastic process that makes the scheme optimal". Muth solved the problem using classical (nonrecursive) methods. The Kalman filter was first described in print in the same year as Muth's solution of this problem. The Kalman filter allows us to solve Muth's problem quickly.

Muth studied the model

$$x_{t+1} = x_t + w_{t+1}, \tag{51}$$
$$y_t = x_t + v_t, \tag{52}$$

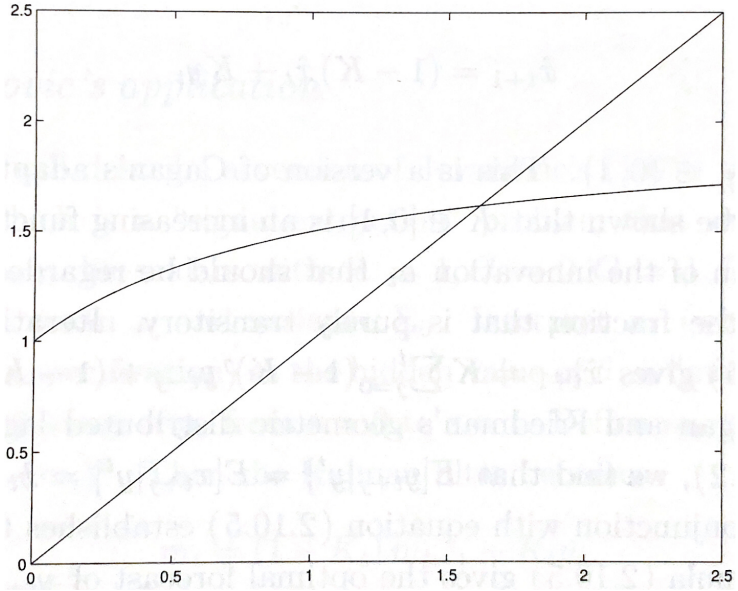
where y_t, x_t are scalar random processes, and w_{t+1}, v_t are mutual independent IID Gaussian random processes with means of 0 and variances $\mathbb{E}[w^w_{t+1}] = Q$, $\mathbb{E}[v^2_t] = R$, and $\mathbb{E}[v_s w_{t+1}] = 0$, $\forall t, s$. The initial condition is that x_0 is Gaussian with mean $\hat{x}_0$ and variance σ^2_0. Muth sought formulas for $\hat{x}_{t+1} = \mathbb{E}[x_{t+1}|y^t]$, where $y^t = [y_t, ..., y_0]$.

For this problem, $\mathbf{A} = 1$, $\mathbf{C}\mathbf{C}^\top = Q$, and $\mathbf{G} = 1$, making the Kalman filtering equations (from (44)

and (47)) become:

$$\begin{aligned}
\mathbf{K}_t &= \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}, \\
\implies K_t &= \frac{\sigma_t^2}{\sigma_t^2 + R}, \qquad (53) \\
\boldsymbol{\Sigma}_{t+1} &= \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top - \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{A}^\top, \\
\implies \sigma_{t+1}^2 &= \sigma_t^2 + Q - \frac{\sigma_t^4}{\sigma_t^2 + R} \\
&= \frac{\sigma_t^2(R+Q) + QR}{\sigma_t^2 + R}. \qquad (54)
\end{aligned}$$

Figure 1: Graph of $f(\sigma^2)$



For $Q = R = 1$, Figure 1 plots the function $f(\sigma^2) = \frac{\sigma^2(R+Q)+QR}{\sigma^2+R}$ appearing on the RHS of (54) for values $\sigma^2 \geq 0$ against the 45-degree line. Note that $f(0) = Q$. This graph identifies the fixed point of iterations on $f(\sigma^2)$ as the intersection of $f(\cdot)$ and the 45-degree line. That the slope of $f(\cdot)$ is less than unity at the intersection assures us that the iterations on f will converge as $t \to +\infty$ starting from any $\sigma_0^2 \geq 0$.

Muth studied the solution of this problem as $t \to \infty$. Evidently, $\sigma_t^2 \to \sigma_\infty^2 \equiv \sigma^2$ is the fixed point

of a graph like Figure 1. Then, $K_t \to K$ and the formula for $\hat{x}_{t+1}$ becomes

$$\hat{x}_{t+1} = (1-K)\hat{x}_t + Ky_t, \tag{55}$$

where $K = \frac{\sigma^2}{\sigma^2+R} \in (0,1)$. This is a version of Cogan's adaptive expectations formula. It can be shown that $K \in [0,1]$ is an increasing function of $\frac{Q}{R}$. Thus, K is the fraction of the innovation a_t that should be regarded as 'permanent' and $1-K$ is the fraction that is purely transitory. Iterating backward on equation (55) gives

$$\hat{x}_{t+1} = K\sum_{j=0}^{t}(1-K)y_{t-j} + (1-K)^{t+1}\hat{x}_0,$$

which is a version of Cagan and Friedman's geometric distributed lag formula.

Using equations (51)-(52), we find that

$$\mathbb{E}[y_{t+j}|y^t] = \mathbb{E}[x_{t+j}|y^t] = \hat{x}_{t+1}, \quad j \geq 1.$$

This result along in conjunction with Equation (55) gives the optimal forecast of y_{t+j} for all horizons $j \geq 1$. This finding is remarkable because for most processes, the optimal forecast will depend on the horizon. That there is a single optimal forecast for all horizons justifies the term **permanent income** that Friedman chose to describe the forecast if income in 1955.

The dependence of the forecast on horizon can be studied using the formulas

$$\mathbb{E}[\mathbf{X}_{t+j}|\mathbf{Y}^{t-1}] = \mathbf{A}^j\hat{\mathbf{X}}_t,$$
$$\mathbb{E}[\mathbf{Y}_{t+j}|\mathbf{Y}^{t-1}] = \mathbf{G}\mathbf{A}^j\hat{\mathbf{X}}_t.$$

In the case of Muth's example,

$$\mathbb{E}[y_{t+j}|y^{t-1}] = \hat{y}_t = \hat{x}_t, \quad \forall j \geq 0.$$

For Muth's model, the innovations representation is

$$\hat{x}_{t+1} = \hat{x}_t + Ka_t,$$
$$y_t = \hat{x}_t + a_t,$$

where $a_t = y_t - \mathbb{E}[y_t|y_{t-1}, y_{t-2}, ...]$. The innovations representation implies that

$$y_{t+1} - y_t = a_{t+1} + (K-1)a_t. \tag{56}$$

This equation represents $\{y_t\}$ as a process whose first difference is a first-order moving average process.

Notice how Friedman's adaptive expectations coefficient, K, appears in this representation.

5.7.2 Example: Jovanovic's matching model

This is from a 1979 paper by Jovanovic. Let x_t, y_t be scalars with $A = 1, C = 0, G = 1, R > 0$. Let $x_0 \sim \mathcal{N}(0, \sigma_0^2)$. Interpret x_t (which is evidently constant with this specification) as the hidden value of θ, a "match parameter". Let y^t denote the history of y_s, from $s = 0$ to $s = t$. Define

$$m_t \equiv \hat{x}_{t+1} \equiv \mathbb{E}[\theta | y^t],$$

and

$$\sigma_{t+1}^2 = \mathbb{E}[\theta - m_t]^2.$$

Then, the Kalman filter becomes

$$m_t = (1 - K_t)m_{t-1} + K_t y_t, \tag{57}$$

$$K_t = \frac{\sigma_t^2}{\sigma_t^2 + R}, \tag{58}$$

$$\sigma_{t+1}^2 = \frac{\sigma_t^2 R}{\sigma_t^2 + R}. \tag{59}$$

The recursions are to be initiated from (m_{-1}, σ_0^2), a pair that embodies all "prior" knowledge about the position of the system. It is easy to see from Figure 1 that when $\mathbf{C}\mathbf{C}^\top = Q = 0, \sigma^2 = 0$ is the limit point of iterations on Equation (59) starting from any $\sigma_0^2 \geq 0$. Thus, the value of the match parameter is eventually learned.

It is instructive to write Equation (59) as

$$\frac{1}{\sigma_{t+1}^2} = \frac{1}{\sigma_t^2} + \frac{1}{R}.$$

The reciprocal of the variance is often called the precision of the estimate. According to this equation, the precision increases without bound as t grows, and $\sigma_{t+1}^2 \to 0$.

We can represent the Kalman filter in the form

$$m_{t+1} = m_t + K_{t+1} a_{t+1},$$

which implies that

$$\mathbb{E}[m_{t+1} - m_t]^2 = K_{t+1}^2 \sigma_{a,t+1}^2,$$

where $a_{t+1} = y_{t+1} - m_t$ and the variance of a_t is equal to $\sigma_{a,t+1}^2 = \sigma_{t+1}^2 + R$. This is because if we

subtract $y_{t+1} = m_t + a_{t+1}$ from $y_{t+1} = x_{t+1} + v_{t+1}$ we get

$$\begin{aligned} a_{t+1} &= x_{t+1} - m_t + v_{t+1} \\ &= x_{t+1} - \hat{x}_{t+1} + v_{t+1}. \end{aligned}$$

The variance of this term is $\sigma_{t+1}^2 + R$. This implies

$$\mathbb{E}[m_{t+1} - m_t]^2 = \frac{\sigma_{t+1}^4}{\sigma_{t+1}^2 + R}.$$

For convenience, we represent the law of motion of m_{t+1} by

$$m_{t+1} = m_t + g_{t+1} u_{t+1},$$

where $g_{t+1} = \left(\frac{\sigma_{t+1}^4}{\sigma_{t+1}^2+R}\right)^{1/2}$ and u_{t+1} is IID normal with mean zero and variance 1 constructed to obey $g_{t+1}u_{t+1} \equiv K_{t+1}a_{t+1}$.

5.8 Example: The LQ permanent income model

To review some key concepts covered (not just the Kalman filter) in this section, we now cover the linear quadratic (LQ) savings problem whose solution is a rational expectations version of the permanent income model of Friedman (1956) and Hall (1978).

The LQ permanent income model is a modification of the following savings problem. A consumer has preferences over consumption streams that are ordered by the utility functional

$$\mathbb{E}_t \sum_{s=0}^{\infty} \beta^s u(c_{t+s}), \tag{60}$$

where $\mathbb{E}_t$ is the expectation operator conditioned on the consumer's time t information, c_t is period t consumption, $u(c_t)$ is a strictly concave one-period utility function, and $\beta \in (0,1)$ is the household discount factor. The consumer maximises utility by choosing consumption and a borrowing plan, $\{c_{t+s}, b_{t+1+s}\}_{s=0}^{\infty}$ subject to the sequence of budget constraints

$$c_{t+s} + b_{t+s} = R^{-1} b_{t+1+s} + y_{t+s}, \tag{61}$$

where y_t is an exogenous stationary endowment process, R is a constant gross risk-free interest rate, b_t is a one-period risk-free debt maturing at t, and b_0 is a given initial condition. We assume that $R^{-1} = \beta$.

For example, we might assume that the endowment process has the state-space representation

$$\mathbf{z}_{t+1} = \mathbf{A}_{22}\mathbf{z}_t + \mathbf{C}_2\mathbf{w}_{t+1}, \tag{62}$$

$$y_t = \mathbf{U}_y\mathbf{z}_t, \tag{63}$$

where w_{t+1} is an IID process with mean zero and identity contemporaneous covariance matrix, $\mathbf{A}_{22}$ is a stable matrix, its eigenvalues within unit circle, and $\mathbf{U}_y$ is a selection vector that identities y with a particular linear combination of the $\mathbf{z}$. We impose the following condition on the consumption and borrowing plan:

$$\mathbb{E}_t \sum_{s=0}^{\infty} \beta^s b_{t+s}^2 < +\infty. \tag{64}$$

This condition suffices to rule out Ponzi schemes. The state vector confronting the household at t is $\begin{bmatrix} b_t \\ \mathbf{z}_t \end{bmatrix}$, where b_t is its one-period debt falling due at the beginning of period t and $\mathbf{z}_t$ contains all variables useful for forecasting its future endowment. The FOCs for maximising (60) subject to (61) are

$$\mathbb{E}_t u'(c_{t+1}) = u'(c_t). \tag{65}$$

For the rest of this section, we assume the quadratic utility function

$$u(c_t) = -\frac{1}{2}(c_t - \gamma)^2,$$

where γ is a bliss level of consumption. Then (65) implies

$$\mathbb{E}_t[c_{t+1}] = c_t. \tag{66}$$

Note that a linear marginal utility is essential for deriving (66) from (65). Suppose instead that we had imposed the following more standard assumptions on the utility function: $u'(c) > 0, u''(c) < 0, u'''(c) > 0$, and $c \geq 0$, like say with log utility. The Euler equation remains the same, but the fact that $u''' < 0$ implies via Jensen's inequality that $\mathbb{E}_t u'(c_{t+1}) > u'(\mathbb{E}_t c_{t+1})$. This inequality together with (65) implies $\mathbb{E}_t c_{t+1} > c_t$ (consumption is said to be a 'submartingale'), so that consumption stochastically diverges to $+\infty$. The consumer's savings also diverge to $+\infty$.

Along with the quadratic utility specification, we allow consumption to be negative (hence why we have the no-Ponzi condition) here.

To deduce the optimal decision rule, we have to solve the system of difference equations formed by (61) and (66) subject to the boundary condition, (64). Solve the period budget constraint, (61),

forward and impose $\lim_{T\to+\infty}\beta^T b_{T+1}=0$ to get

$$\begin{aligned}
c_t+b_t &= R^{-1}b_{t+1}+y_t,\\
c_{t+1}+b_{t+1} &= R^{-1}b_{t+2}+y_{t+1},\\
&\vdots\\
\implies c_t+b_t &= R^{-1}\left(R^{-1}b_{t+2}+y_{t+1}-c_{t+1}\right)+y_t\\
b_t &= R^{-1}R^{-1}b_{t+2}+y_t+R^{-1}y_{t+1}-(c_t+R^{-1}c_{t+1})\\
&\vdots
\end{aligned}$$

and with some cleaning up:

$$b_t=\sum_{j=0}^{\infty}\beta^j(y_{t+j}-c_{t+j}). \tag{67}$$

Imposing $\lim_{T\to+\infty}\beta^T b_{T+1}=0$ suffices to impose (64) on the debt path. Take conditional expectations on both sides of (67) and use (66) and the LIE to deduce

$$b_t=\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j}+\frac{1}{1-\beta}c_t \tag{68}$$

$$\Leftrightarrow c_t=(1-\beta)\left[\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j}-b_t\right]. \tag{69}$$

If we define the net interest rate, r, by $\beta=R^{-1}=\frac{1}{1+r}$, we can write the above expression as

$$c_t=\frac{r}{1+r}\left[\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j}-b_t\right]. \tag{70}$$

Equation (69) or (70) expresses consumption as equaling economic income, namely, a constant marginal propensity to consume or interest factor, $\frac{r}{1+r}$, times the sum of nonfinancial wealth and financial wealth. Note also that these expressions represents c_t as a function of the states confronting the household, where from (62)-(63) $\mathbf{z}_t$ contains the information useful for forecasting the endowment process.

5.8.1 Another representation

Pulling together our preceding results, we can regard $\mathbf{z}_t, b_t$ as the time t states, where $\mathbf{z}_t$ are the exogenous components and b_t is the endogenous component of the state vector. The system can be

represented as

$$\begin{aligned}
\mathbf{z}_{t+1} &= \mathbf{A}_{22}\mathbf{z}_t + \mathbf{C}_2\mathbf{w}_{t+1}, \\
b_{t+1} &= b_t + \mathbf{U}_y\left[(\mathbf{I}-\beta\mathbf{A}_{22})^{-1}(\mathbf{A}_{22}-\mathbf{I})\right]\mathbf{z}_t, \\
y_t &= \mathbf{U}_y\mathbf{z}_t, \\
c_t &= (1-\beta)\left[\mathbf{U}_y(\mathbf{I}-\beta\mathbf{A}_{22})^{-1}\mathbf{z}_t - b_t\right].
\end{aligned}$$

Another way to understand the solution is to show that after the optimal decision rule has been obtained, there is a point of view that allows us to regard the state as being c_t together with $\mathbf{z}_t$ and to regard $\mathbf{b}_t$ as the outcome. Following Hall (1978), this is a sharp way to summarise the implication of the LQ permanent income theory. We now proceed to transform the state vector in this way.

To represent the solution for b_t, substitute (69) into (61) to get

$$\begin{aligned}
(1-\beta)\left[\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j} - b_t\right] + b_t &= R^{-1}b_{t+1} + y_t \\
R^{-1}b_{t+1} &= (1-\beta)\left[\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j} - b_t\right] + b_t - y_t \\
b_{t+1} &= R(1-\beta)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j} + (R - R(1-\beta))b_t - Ry_t \\
b_{t+1} &= b_t + (\beta^{-1}-1)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j} - \beta^{-1}y_t. \qquad (71)
\end{aligned}$$

Next, shift (69) forward one period and eliminate b_{t+1} by using (61) to obtain:

$$c_{t+1} = (1-\beta)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_{t+1}y_{t+1+j} - (1-\beta)\left[\beta^{-1}(c_t + b_t - y_t)\right],$$

and if we add and subtract $\beta^{-1}(1-\beta)\sum_{j=0}^{\infty}\beta^j\mathbb{E}_t y_{t+j}$ from the RHS and rearrange, we get

$$c_{t+1} - c_t = (1-\beta)\sum_{j=0}^{\infty}\beta^j\left(\mathbb{E}_{t+1}y_{t+j+1} - \mathbb{E}_t y_{t+j+1}\right). \qquad (72)$$

The RHS is the time $t+1$ innovation to the expected present value of the endowment process, y. It is useful to express this invariation in terms of a moving average (MA) representation for income, y_t. Suppose that the endowment process has the MA representation

$$y_{t+1} = \mathbf{d}(L)\mathbf{w}_{t+1}, \qquad (73)$$

where $\mathbf{w}_{t+1}$ is an IID vector process with $\mathbb{E}_t \mathbf{w}_{t+1} = \mathbf{0}$, and contemporaneous covariance matrix $\mathbb{E}_t \mathbf{w}_{t+1}\mathbf{w}_{t+1}^\top = \mathbf{I}$, $\mathbf{d}(L) = \sum_{j=0}^{\infty} \mathbf{d}_j L^j$, where L is the lag operator, and the household has an information set $\mathbf{w}^t = [\mathbf{w}_t, \mathbf{w}_{t-1}, , ...]$ at time t. Then notice that

$$y_{t+j} - \mathbb{E}_t y_{t+j} = \mathbf{d}_0 \mathbf{w}_{t+j} + \mathbf{d}_1 \mathbf{w}_{t+j-1} + \cdots + \mathbf{d}_{j-1}\mathbf{w}_{t+1}.$$

It follows that

$$\mathbb{E}_{t+1} y_{t+j} - \mathbb{E}_t y_{t+j} = \mathbf{d}_{j-1}\mathbf{w}_{t+1}. \tag{74}$$

Using (74) in (72) gives

$$c_{t+1} - c_t = (1-\beta)\mathbf{d}(\beta)\mathbf{w}_{t+1}. \tag{75}$$

The object $\mathbf{d}(\beta)$ is the present value of the moving average coefficients in the representation for the endowment process y_t.

After all of this work, we can represent the optimal decision rule for c_t, b_{t+1} in the form of the two equations, (68) and (72), which for reference are:

$$b_t = \sum_{j=0}^{\infty} \beta^j \mathbb{E}_t y_{t+j} + \frac{1}{1-\beta} c_t,$$

$$c_{t+1} - c_t = (1-\beta) \sum_{j=0}^{\infty} \beta^j \left(\mathbb{E}_{t+1} y_{t+j+1} - \mathbb{E}_t y_{t+j+1} \right).$$

Equation (68) assets that the household's debt due at t equals the expected present value of its endowment minus the expected present value of its consumption stream. A high debt thus indicates a large expected present value of 'surpluses' $y_t - c_t$.

Recalling the form of the endowment process (63), we can compute

$$\mathbb{E}_t \sum_{j=0}^{\infty} \beta^j \mathbf{z}_{t+j} = (\mathbf{I} - \beta \mathbf{A}_{22})^{-1} \mathbf{z}_t$$

$$\mathbb{E}_{t+1} \sum_{j=0}^{\infty} \beta^j \mathbf{z}_{t+j+1} = (\mathbf{I} - \beta \mathbf{A}_{22})^{-1} \mathbf{z}_{t+1}$$

$$\mathbb{E}_t \sum_{j=0}^{\infty} \beta^j \mathbf{z}_{t+j+1} = (\mathbf{I} - \beta \mathbf{A}_{22})^{-1} \mathbf{A}_{22} \mathbf{z}_t.$$

Substituting these formulas into (68) and (72), and using (62), gives the following representation for

the consumer's optimum decision rule:

$$c_{t+1} = c_t + (1-\beta)\mathbf{U}_y(\mathbf{I} - \beta\mathbf{A}_{22})^{-1}\mathbf{C}_2\mathbf{w}_{t+1}, \tag{76}$$

$$b_t = \mathbf{U}_y(\mathbf{I} - \beta\mathbf{A}_{22})^{-1}\mathbf{z}_t - \frac{1}{1-\beta}c_t, \tag{77}$$

$$y_t = \mathbf{U}_y\mathbf{z}_t, \tag{78}$$

$$\mathbf{z}_{t+1} = \mathbf{A}_{22}\mathbf{z}_t + \mathbf{C}_2\mathbf{w}_{t+1}. \tag{79}$$

The above representation reveals several things about the optimal decision rule.

1. The state consists of the endogenous part, c_t, and the exogenous part, $\mathbf{z}_t$. These contain all of the relevant information for forecasting future c, y, b. Notice that financial assets, b_t, have disappeared as a component of the state because they are properly encoded in c_t.

2. Consumption is a random walk with innovation $(1-\beta)\mathbf{d}(\beta)\mathbf{w}_{t+1}$ as implied also by (75). This outcome confirms that the Euler equation (66) is built into the solution. That consumption is a random walk of course implies that it does not possess an asymptotic stationary distribution, at least so long as $\mathbf{z}_t$ exhibits perpetual random fluctuations, as it will generally under (62). This feature is inherited partly from the assumption that $\beta R = 1$. The failure of consumption to converge is something to consider when we drop quadratic utility and assume consumption must be nonnegative.

3. The impulse response function of c_t is a box: for all $j \geq 1$, the response of c_{t+j} to an increase in the innovation $\mathbf{w}_{t+1}$ is

$$(1-\beta)\mathbf{d}(\beta) = (1-\beta)\mathbf{U}_y(1 - \beta\mathbf{A}_{22})^{-1}\mathbf{C}_2.$$

4. Solution (76)-(79) reveals that the joint process c_t, b_t possess the property that Engle and Granger (1987) called **cointegration**. In particular, both c_t and b_t possess a unit-root (see (71) for b_t), but there is a linear combination of c_t, b_t that is stationary provided that $\mathbf{z}_t$ is stationary. From (68), the linear combination is

$$(1-\beta)b_t + c_t.$$

Accordingly, Engle and Granger would call $\begin{bmatrix}(1-\beta) & 1\end{bmatrix}$ a cointegrating vector that, when applied to the nonstationary vector process, $\begin{bmatrix}b_t \\ c_t\end{bmatrix}$, yields a process that is asymptotically stationary. Equation (68) can be arranged to take the form

$$(1-\beta)b_t + c_t = (1-\beta)\mathbb{E}_t\sum_{j=0}^{\infty}\beta^j y_{t+j},$$

which assets that the 'cointegrating residual' on the LHS equals the conditional expectation of the geometric sum of future incomes on the RHS.

5.8.2 Debt dynamics

If we subtract Equation (77) evaluated at time t from Equation (77) evaluated at time $t+1$ we obtain

$$b_{t+1} - b_t = \mathbf{U}_y(1 - \beta\mathbf{A}_{22})^{-1}(\mathbf{z}_{t+1} - \mathbf{z}_t) - \frac{1}{1-\beta}(c_{t+1} - c_t).$$

Substituting $\mathbf{z}_{t+1} - \mathbf{z}_t = (\mathbf{A}_{22} - \mathbf{I})\mathbf{z}_t + \mathbf{C}_2\mathbf{w}_{t+1}$ and equation (76) into the above equation and rearranging gives

$$b_{t+1} - b_t = \mathbf{U}_y(\mathbf{I} - \beta\mathbf{A}_{22})^{-1}(\mathbf{A}_{22} - \mathbf{I})\mathbf{z}_t.$$

6 Maximum Likelihood

The ML method is one of the most popular ways to estimate the parameter θ that specifies a probability function $\Pr(X = x|\theta)$ of a discrete stochastic variable X (or a probability density function $\phi(x|\theta)$ of a continuous stochastic variable X) based on the observations $x_1, ..., x_n$ which are independently sampled from the distribution.

Again, we won't go a full description here. The main goal is to cover the basics and then link up what we did with the Kalman filter to ML estimation.

6.1 Estimation

Unlike the GMM approach, the ML method requires one to know the full distribution of the DGP. Suppose that the observable data, $\{x_1, ..., x_T\}$ are independently and identically drawn from a PDF $\phi(\cdot, \boldsymbol{\theta})$ given a parameter $\boldsymbol{\theta}$. The joint distribution is given by

$$\phi(x_1, ..., x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \phi(x_i, \boldsymbol{\theta}).$$

The ML method is designed to maximise the likelihood function for the entire sample:

$$L(\boldsymbol{\theta}|x_1, ..., x_n) = \phi(x_1, ..., x_n|\boldsymbol{\theta}).$$

In practice, it's easier to work with the log of the likelihood function:

$$\max_{\boldsymbol{\theta}} \ \ln L(\boldsymbol{\theta}|x_1, ..., x_n) = \sum_{i=1}^{n} l(x_i, \boldsymbol{\theta}).$$

6.1.1 Example: Search and match

Consider a simple job search problem. Suppose that job offers are independently and identically drawn from a fixed known distribution, F. The Bellman equation is given by

$$V(w) = \max\left\{\frac{w}{1-\beta}, c + \beta\int V(w')dF(w')\right\}.$$

There is a cutoff value, $w^*(\boldsymbol{\theta})$, such that the worker takes the job offer, w, if and only if $w \geq w^*(\theta)$, where $\boldsymbol{\theta}$ represents the parameters c (unemployment compensation) and β. The reservation wage, $w^*(\boldsymbol{\theta})$, is unobservable. But we can compute it numerically given any parameter value $\boldsymbol{\theta}$. We can then compute the likelihood of observing a worker i accepting a job for the first time after t_i periods:

$$L_i(\boldsymbol{\theta}) = (1 - F(w^*(\boldsymbol{\theta})))\left[F(w^*(\boldsymbol{\theta}))\right]^{t_i - 1}.$$

Say we observe durations t_i for n workers. Then the likelihood of the sample is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}).$$

6.2 Asymptotic properties

Suppose that ϕ is differentiable and concave in $\boldsymbol{\theta}$. The FOC for the log-likelihood function is given by

$$\sum_{i=1}^{n} \frac{\partial l(x_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

The population analogue is given by

$$\mathbb{E}\left[\frac{\partial l(x_i, \theta)}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

This moment condition implies the ML estimator can be viewed as a GMM estimator with $\mathbf{h}(x_t, \boldsymbol{\theta}) = \partial l(x_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$. We can then apply the results derived in the previous sections. We list these properties without explicitly stating relevant conditions and proofs.

Consistency is given by

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta},$$

and the ML estimator is asymptotically normally distributed

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(0, \mathbf{I}(\boldsymbol{\theta})^{-1}\right),$$

where

$$\mathbf{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} l(\boldsymbol{\theta})\right],$$

which is defined as the Fisher information matrix. By the information matrix equality,

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} l(\boldsymbol{\theta})\right] \\ &= \mathbb{E}\left[\left(\frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^\top \frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right].\end{aligned}$$

The ML estimator is asymptotically efficient: it achieves the Cramer-Rao lower bound when the sample size tends to infinity. This means that no asymptotically unbiased estimator has lower asymptotic mean-squared error than the ML estimator.

6.3 Back to the Kalman filter

Recall our latent variable problem, (28)-(29):

$$\begin{aligned}\mathbf{X}_{t+1} &= \mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1}, \\ \mathbf{Y}_t &= \mathbf{G}\mathbf{X}_t + \mathbf{V}_t,\end{aligned}$$

where $\mathbf{X}$ was the unobserved n-vector of states and $\mathbf{Y}$ was its m-vector of signals which we observed, and $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We derived the Kalman filter, (43)-(47)

$$\begin{aligned}\mathbf{a}_t &= \mathbf{Y}_t - \mathbf{G}\hat{\mathbf{X}}_t, \\ \mathbf{K}_t &= \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}, \\ \hat{\mathbf{X}}_{t+1} &= \mathbf{A}\hat{\mathbf{X}}_t + \mathbf{K}_t\mathbf{a}_t, \\ \boldsymbol{\Sigma}_t &= \mathbf{C}\mathbf{C}^\top + \mathbf{K}_t\mathbf{R}\mathbf{K}_t^\top + (\mathbf{A} - \mathbf{K}_t\mathbf{G})\boldsymbol{\Sigma}_t(\mathbf{A} - \mathbf{K}_t\mathbf{G})^\top,\end{aligned}$$

and with

$$\boldsymbol{\Sigma}_{t+1} = \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top - \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{G}^\top(\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{G}^\top + \mathbf{R})^{-1}\mathbf{G}\boldsymbol{\Sigma}_t\mathbf{A}^\top.$$

We briefly discussed estimation, but let's focus on that in this section. Given data on $\mathbf{Y}_t$, and some initial conditions, we can use ML to estimate $\boldsymbol{\Psi} = (\mathbf{A}, \mathbf{G}, \mathbf{C}, \mathbf{R})$.

The innovations representation that emerges from the Kalman filter is

$$\begin{aligned}\hat{\mathbf{X}}_{t+1} &= \mathbf{A}\hat{\mathbf{X}}_t + \mathbf{K}_t\mathbf{a}_t, \\ \mathbf{Y}_t &= \mathbf{G}\hat{\mathbf{X}}_t + \mathbf{a}_t,\end{aligned}$$

where for $t \geq 1$, $\hat{\mathbf{X}}_t = \mathbb{E}[\mathbf{X}_t|\mathbf{Y}^{t-1}]$ and $\mathbb{E}[\mathbf{a}_t\mathbf{a}_t^\top] = \mathbf{G}\Sigma_t\mathbf{G}^\top + \mathbf{R} = \mathbf{\Omega}_t$, so we have $\mathbf{a}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)$. Evidently, for $t \geq 1$, $\mathbb{E}[\mathbf{Y}_t|\mathbf{Y}^{t-1}] = \mathbf{G}\hat{\mathbf{X}}_t$, and the so

$$\mathbf{Y}_t|\mathbf{Y}^{t-1} \sim \mathcal{N}(\mathbf{G}\hat{\mathbf{X}}_t, \mathbf{\Omega}_t).$$

The objects $\mathbf{G}\hat{\mathbf{X}}_t, \mathbf{\Omega}_t$ emerging from the Kalman filter are thus sufficient statistics and also the innovation $\mathbf{a}_t = \mathbf{Y}_t - \mathbf{G}\hat{\mathbf{X}}_t$ can be calculated recursively from (43)-(46).

We can factor the likelihood function for a sample $(\mathbf{Y}_T, \mathbf{Y}_{T-1}, ..., \mathbf{Y}_0)$ as

$$\phi(\mathbf{Y}_T, ..., \mathbf{Y}_0) = \phi(\mathbf{Y}_T|\mathbf{Y}^{T-1})\phi(\mathbf{Y}_{T-1}|\mathbf{Y}^{T-2}) \cdots \phi(\mathbf{Y}_1|\mathbf{Y}_0)\phi(\mathbf{Y}_0). \tag{80}$$

The log of the conditional density of the $m \times 1$ vector $\mathbf{Y}_t$ is

$$\log \phi(\mathbf{Y}_t|\mathbf{Y}^{t-1}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Omega}_t| - \frac{1}{2}\mathbf{a}_t^\top \mathbf{\Omega}_t \mathbf{a}_t. \tag{81}$$

We can use (81) and (43)-(46) to evaluate the likelihood function (80) recursively for a given set of parameter values $\mathbf{\Psi}$ that underlie the matrices $\mathbf{A}, \mathbf{G}, \mathbf{C}, \mathbf{R}$. Such calculations are at the heart of efficient strategies for computing MLEs of free parameters. For example, suppose we observe a sequence of data, $\mathbf{Y}^T$, generated by the state space system (28) and (29). Its log-likelihood density is

$$\begin{aligned}\log \phi(\mathbf{Y}^T|\mathbf{\Psi}) &= \sum_{t=1}^{T} \ln(\mathbf{Y}_t|\mathbf{Y}^{t-1}) \\ &= -\frac{mT}{2}\log(2\pi) - \sum_{t=1}^{T}\frac{1}{2}\log|\mathbf{\Omega}_t| - \frac{1}{2}\sum_{t=1}^{T}\mathbf{a}_t^\top \mathbf{\Omega}_t \mathbf{a}_t.\end{aligned}$$

The likelihood function is also an essential object for Bayesian statistics.[18] It completely summarises how the data influence the Bayesian posterior via the following application of Bayes' Law. Where $\boldsymbol{\theta}$ is our parameter vector, $\mathbf{Y}_0^T$ our data record, and $\tilde{p}(\boldsymbol{\theta})$ a probability density that summarises our prior 'views' or 'information' about $\boldsymbol{\theta}$ before seeing $\mathbf{Y}_0^T$, our views about $\boldsymbol{\theta}$ after seeing $\mathbf{Y}_0^T$ are described by a posterior probability, $\tilde{p}(\boldsymbol{\theta}|\mathbf{Y}_0^T)$ that is constructed from Bayes' Law via

$$\tilde{p}(\boldsymbol{\theta}|\mathbf{Y}_0^T) = \frac{\phi(\mathbf{Y}_0^T|\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta})}{\int \phi(\mathbf{Y}_0^T|\boldsymbol{\theta})\tilde{p}(\boldsymbol{\theta})d\boldsymbol{\theta}},$$

where the denominator is the marginal joint density, $\phi(\mathbf{Y}_0^T)$, of $\mathbf{Y}_0^T$.

[18] See for example DeJong and Dave (2012) and Fernández-Villaverde et al. (2016).

6.4 Back to DSGE models

6.4.1 Example: Baseline RBC model

Consider the linearised RBC model:

$$
\begin{aligned}
y_t &= \left(1 - \frac{\alpha\delta}{\beta^{-1} + \delta - 1}\right) c_t + \left(\frac{\alpha\delta}{\beta^{-1} + \delta - 1}\right) i_t, \\
y_t &= a_t + \alpha k_{t-1} + (1-\alpha) n_t, \\
k_t &= \delta i_t + (1-\delta) k_{t-1}, \\
n_t &= y_t - \eta c_t, \\
c_t &= \mathbb{E}_t c_{t+1} - \frac{1}{\eta} \mathbb{E}_t r_{t+1}, \\
r_t &= (1 - \beta(1-\delta))(y_t - k_{t-1}), \\
a_t &= \rho a_{t-1} + \epsilon_t.
\end{aligned}
$$

This model features seven equations in six endogenous variables, $y_t, c_t, i_t, k_t, n_t, r_t$, and one exogenous variable, a_t. The challenge here – and in most DSGE models – is that we can only observe y_t, c_t, i_t, and n_t (or at least the HP-filtered version of them that we are likely to use to estimate the model). But we don't observe a_t and since we don't really know depreciation rates, this means we don't observe k_t or n_t. So this model mixes four observable variables with three unobservable variables.

Models like the RBC model provide a micro-foundation for why we cannot find a perfect fitting model with the observed data: There is an unobservable technology series and all of the observed series depend on this. However, it is still not possible to estimate this joint model by ML techniques. This is because the same unobserved series shows up in all of the reduced-form solution equations. So while the model features stochastic shocks, it has a feature that is known as a stochastic singularity: The shocks in all of the equations are just multiples of each other.

The model thus predicts that certain ratios of the observed variables (e.g. current and lagged consumption, current and lagged investment) will be constant. In practice, these prediction will not hold in the data, so there is no chance that this model can fit the data.

In general, for a model to have well–defined econometric estimates, it is necessary that **for every observable variable there be at least one unobservable shock**. This can either take the form of a "measurement error" or else involve a shock in each equation with a clear structural interpretation.

Log-linearised DSGE models with a mix of observable and unobservable variables are an example of state-space models – something that we should be familiar with having covered the Kalman filter. Recall that these models can be described using two equations. The first, known as the state or

transition equation, describes how a set of unobservable state variables, $\mathbf{X}_t$, evolve over time:

$$\mathbf{X}_{t+1} = \mathbf{A}\mathbf{X}_t + \mathbf{C}\mathbf{W}_{t+1}.$$

The second equation in a state-space model, which is known as the measurement equation, relates a set of observable signals, $\mathbf{Y}_t$, to the unobservable state variables

$$\mathbf{Y}_t = \mathbf{G}\mathbf{X}_t + \mathbf{V}_t.$$

The solution to the baseline RBC model without labour input can be summarised as

$$\begin{aligned} k_t &= a_{kk}k_{t-1} + a_{kz}z_t, \\ c_t &= a_{ck}k_{t-1} + a_{cz}z_t, \\ z_t &= \rho z_{t-1} + \epsilon_t. \end{aligned}$$

This output is something that Dynare will give you for a set of parameters (just be careful with the timings). Now, let's assume that consumption and capital are only observed with error so that the two observable variables are

$$\begin{aligned} k_t^* &= a_{kk}k_{t-1} + a_{kz}z_t + v_t^k, \\ c_t^* &= a_{ck}k_{t-1} + a_{cz}z_t + v_t^c. \end{aligned}$$

The transition equation is:

$$\underbrace{\begin{bmatrix} k_t \\ z_{t+1} \end{bmatrix}}_{\mathbf{X}_{t+1}} = \underbrace{\begin{bmatrix} a_{kk} & a_{kz} \\ 0 & \rho \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} k_{t-1} \\ z_t \end{bmatrix}}_{\mathbf{X}_t} + \underbrace{\begin{bmatrix} 0 \\ \epsilon_{t+1} \end{bmatrix}}_{\mathbf{W}_{t+1}},$$

where we simply assume $\mathbf{C} = \mathbf{I}$. The measurement equation is

$$\underbrace{\begin{bmatrix} k_{t-1}^* \\ c_t \end{bmatrix}}_{\mathbf{Y}_t} = \underbrace{\begin{bmatrix} 1 & 0 \\ a_{ck} & a_{cz} \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} k_{t-1} \\ z_t \end{bmatrix}}_{\mathbf{X}_t} + \underbrace{\begin{bmatrix} v_{t-1}^k \\ v_t^c \end{bmatrix}}_{\mathbf{V}_t},$$

and notice that we had to do some trickery to get the model in state-space form – this is macroeconomics after all, and there's always a trick. Nevertheless, all standard DSGE models can be re-arranged to be put in this format.

What if we didn't observe capital but instead observed output, y_t^*? Then we would have:

$$\underbrace{\begin{bmatrix} k_t \\ z_{t+1} \end{bmatrix}}_{\mathbf{X}_{t+1}} = \underbrace{\begin{bmatrix} a_{kk} & a_{kz} \\ 0 & \rho \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} k_{t-1} \\ z_t \end{bmatrix}}_{\mathbf{X}_t} + \underbrace{\begin{bmatrix} 0 \\ \epsilon_{t+1} \end{bmatrix}}_{\mathbf{W}_{t+1}},$$

$$\underbrace{\begin{bmatrix} c_t^* \\ y_t^* \end{bmatrix}}_{\mathbf{Y}_t} = \underbrace{\begin{bmatrix} a_{ck} & a_{cz} \\ \alpha \bar{z} \bar{k}^{\alpha-1} & \bar{k}^{\alpha} \end{bmatrix}}_{\mathbf{G}} \underbrace{\begin{bmatrix} k_{t-1} \\ z_t \end{bmatrix}}_{\mathbf{X}_t} + \underbrace{\begin{bmatrix} v_t^c \\ 0 \end{bmatrix}}_{\mathbf{V}_t}.$$

We can then put this into Dynare – or any other software program – and use ML and the Kalman filter to attain our model estimates.

But, there's a catch – as always. If you think that this is a complicated process where things might go wrong, then you'd be right. Read the paper "The Econometrics of DGSE Models" by Fernández-Villaverde (2010). He discusses some of the problems associated with MLE for DSGE models and explains why a Bayesian approach of calculating the full posterior distribution may be preferable:

> "[...]maximising a complicated, highly dimensional function like the likelihood of a DSGE model is actually much harder than it is to integrate it, which is what we do in a Bayesian exercise. First, the likelihood of DSGE models is, as I have just mentioned, a highly dimensional object, with a dozen or so parameters in the simplest cases to close to a hundred in some of the richest models in the literature. Any search in a high dimensional function is fraught with peril. More pointedly, likelihoods of DSGE models are full of local maxima and minima and of nearly flat surfaces. This is due both to the sparsity of the data (quarterly data do not give us the luxury of many observations that micro panels provide) and to the flexibility of DSGE models in generating similar behaviour with relatively different combination of parameter values [...] Moreover, the standard errors of the estimates are notoriously difficult to compute and their asymptotic distribution a poor approximation to the small sample one."

References

Anderson, E. W., McGrattan, E. R., Hansen, L. P., and Sargent, T. J. (1996), "Mechanics of Forming and Estimating Dynamic Linear Economies", *Handbook of Computational Economics*, 1: 171–252.

Cochrane, J. H. (2005), *Asset Pricing (Revised Edition)* (Princeton University Press).

Davidson, R. and MacKinnon, J. G. (2004), *Econometric Theory and Methods* (Oxford University Press).

DeJong, D. N. and Dave, C. (2012), *Structural Macroeconometrics* (2nd Edition, Princeton University Press).

Engle, R. F. and Granger, C. W. J. (1987), "Co-Integration and Error Correction: Representation, Estimation, and Testing", *Econometrica*, 55/2: 251–76.

Fernández-Villaverde, J. (2010), "The Econometrics of DSGE Models", *SERIEs*: 3–49.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., and Schorfheide, F. (2016), "Solution and Estimation Methods for DSGE Models", *Handbook of Macroeconomics*, 2: 527–724.

Hall, R. E. (1978), "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence", *Journal of Political Economy*, 86/6: 971–87.

Hansen, G. D. (1985), "Indivisible Labor and the Business Cycle", *Journal of Monetary Economics*, 16: 309–27.

Hansen, G. D. and Wright, R. (1992), "The Labor Market in Real Business Cycle Theory", *Quarterly Review*, 16: 2–12.

Hansen, L. P. (1982), "Large Sample Properties of Generalised Method of Moments Estimators", *Econometrica*, 50/4: 1029–54.

Hayashi, F. (2000), *Econometrics* (Princeton University Press).

Kydland, F. E. and Prescott, E. C. (1982), "Time to Build and Aggregate Fluctuations", *Econometrica*, 50/6: 1345–70.

Ljungqvist, L. and Sargent, T. J. (2018), *Recursive Macroeconomic Theory* (4th Edition, MIT Press).

McFadden, D. (1987), "Regression-Based Specification Tests for the Mulitnomial Logit Model", *Journal of Econometrics*, 34/1-2: 63–82.

McFadden, D. (1989), "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration", *Econometrica*, 57/5: 995–1026.

Miao, J. (2020), *Economic Dynamics in Discrete Time* (2nd Edition, MIT Press).

Newey, W. K. and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55/3: 703–8.

Pfeifer, J. (2013), "A Guide to Specifying Observation Equations for the Estimation of DSGE Models", *(draft version September 17, 2020)*.

Prescott, E. C. (1986), "Theory Ahead of Business-Cycle Measurement", *Carnegie-Rochester Conference Series on Public Policy*, 25: 11–44.

Romer, D. H. (2012), *Advanced Macroeconomics* (4th Edition, McGraw-Hill Irwin).