Contents

Chapter 4

Maximum Likelihood Estimation and Optimization

4.1 Introduction into Maximum Likelihood Estimation (MLE)

- In this Chapter, we review the MLE discussion in Chapter 1
- Specialize and discuss the estimation of the parameters of time series models by **Maximum Likelihood**.
- After we have outlined the MLE theory, we will consider the numerical optimization problem.

1. The joint density function for observations $y_1, \ldots, y_T$ is assumed to depend upon n unknown parameters in the vector $\Psi = (\psi_1, \ldots, \psi_n)^T$. Note that T is transpose which should be obvious from the context to separate it from the number of observations also denoted by T.

2. Denote the *joint density* in T observations:

$$L(y_1, \ldots y_T; \Psi) \tag{4.1}$$

3. The *likelihood* is a reinterpretation of the density, as a function of the parameters Ψ given the observed or drawn sample . i.e. given the data $y_1 \ldots y_T$, we consider different values for the Ψ

$$L(\Psi; y_1, \ldots, y_T) \text{ or } L(\Psi)$$

4. Object of MLE is to maximize $L(\Psi)$ or more commonly the log likelihood (a monotonic transform of the likelihood): $\log(L)$ w.r.t. to choice over Ψ.

5. The maximum likelihood estimator is obtained by setting the **score** $\frac{\partial logL}{\partial \Psi}$ to zero:
$$\frac{\partial logL}{\partial \Psi} = 0 \qquad \}\text{likelihood equations} \tag{4.2}$$

6. Likelihood equations are typically nonlinear in the parameters which will require an iterative procedure like Gauss-Newton.

7. Even though the models we have dealt with previously have been the fixed X with iid support (or some form of heteroskedasticity or autoregressive errors), all the desirable properties of MLE obtain for these kinds of time series models given our assumptions.

4.1.1 Properties for MLE

Let $\tilde{\Psi} = \{\tilde{\Psi}_1, \ldots, \tilde{\Psi}_n\}$ be the MLE estimate

$$\sqrt{T}(\tilde{\Psi} - \Psi) \underset{\sim}{d} N(0, T\ Avar) \tag{4.3}$$

where **AVar** is the asymptotic variance-covariance matrix obtained in the usual way from the probability limit of the inverse **information matrix (IA)**:

$$\mathbf{IA}(\Psi) = p\lim T^{-1}\{-\frac{\partial^2 log(L)}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T}\}. \tag{4.4}$$

that is

$$\mathbf{Avar}(\tilde{\Psi}) = T^{-1}\mathbf{IA}(\Psi)^{-1} \tag{4.5}$$

We say the parameters from MLE are $\sqrt{T}$ consistent (read as root T conistent). A trvial example from early statistics will illustrae the point nicely

4.2 Maximum Likelihood of μ and σ^2 in $y_t \sim NID(\mu, \sigma^2)$

- Suppose we were interested in estimating μ and σ^2 in the simple case

$$y_t \sim NID(\mu, \sigma^2) \qquad t = 1, \ldots T \tag{4.6}$$
$$y_t = \mu + \epsilon_t \qquad \epsilon_t \sim NID(0, \sigma^2) \tag{4.7}$$

- We know under these conditions that the Ordinary Least Squares estimator ((OLS) is *Best Linear Unbiased Estimator (BLUE)* for μ and σ^2 is

$$\hat{\mu} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{4.8}$$
$$s^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_t - \bar{y})^2$$

- We might ask what are the Maximum Likelihood Estimators of μ and σ^2?
- First we recall the density for a normal variable

$$L(y_t; \mu, \sigma^2) = f(y_t) = (2\pi\sigma^2)^{-\frac{1}{2}} \; e^{\frac{-(y_t-\mu)}{2\sigma^2}} \tag{4.9}$$

- We view the density as the distribution of y_t given the parameters μ and σ^2.
- Now since y_t is *independent* for all T, the **joint density** is the **product** of the marginal densities:

$$L(y_1, \ldots, y_T; \mu, \sigma^2) = f(y_1)f(y_2) \times \cdots \times f(y_T) \tag{4.10}$$

- This may be written as

$$L(y_1, \ldots, y_T; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{T}{2}} e^{\; -\frac{\sum_{t=1}^{T}(y_t-\mu)^2}{2\sigma^2}} \tag{4.11}$$

- Therefore, the **log likelihood** (since we are going to take derivatives of this it is much easier to deal with the logarithms) is obtained by reinterpreting the density as a function of the parameters μ and σ^2 **given the data** and taking logarithms

$$\begin{aligned} \log(L) \;\; = \;\; & \mathcal{L}(\mu, \sigma^2; \mathbf{y}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(2\sigma^2) \\ & -\frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \mu)^2 \end{aligned} \tag{4.12}$$

- The MLE are obtained by taking the first derivative of (4.12) with respect to μ and σ^2 and setting them to zero.
- First

$$\frac{\partial \mathcal{L}(\mu, \sigma^2; \mathbf{y})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{t=1}^{T}(y_t - \mu) \tag{4.13}$$

- Setting this to zero and solving for $\tilde{\mu}$ gives

$$\tilde{\mu} = \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{4.14}$$

  which shows that the MLE for this simple model is *identical to the sample mean*!

- Next take the derivative of the log likelihood with respect to σ^2

$$\frac{\partial \mathcal{L}(\mu, \sigma^2; \mathbf{y})}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{T}(y_t - \mu)^2 \tag{4.15}$$

- Setting this to zero and solving for $\tilde{\sigma}^2$ (substituting $\tilde{\mu}$, the MLE of μ) gives:

$$\tilde{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \tilde{\mu})^2 \tag{4.16}$$

- Notice carefully that this is not the same as s^2 since we are dividing by T and not $(T-1)$.

- To see the limiting distribution we note first that $V[\overline{Y}] = \frac{\sigma^2}{T} \to 0$ as $T \to \infty$ so $\bar{Y}$ has a degenerative distribution (its consistent) to get a limiting distribution we must "slow" down the collapse of the distribution by multiplying by $\sqrt{T}$

$$\sqrt{T}(\bar{Y} - \mu) \sim N(0, \sigma^2)$$

- In general, MLE are biased but **consistent** (recall consistency property from Chapter 3).

- In practice, we evaluate the **Avar** at the MLE : $\tilde{\Psi}$. This is the estimates of the **Avar** produced by the program

4.2.1 Notes on MLE and Regression

1. Review maximum likelihood estimation in the classical linear regression model. There are many good references for this (Davidson and MacKinnon or Greene will do).

2. With an autoregressive error terms u in

$$y = X\beta + u$$

the estimation procedure in the generalized regression model is generalized least squares (GLS). The same basic principle is followed when the error structure is MA.

3. Suppose we have normal support $u \sim N(\mathbf{0}, \Omega)$, then the likelihood is:

$$logL = -\frac{T}{2}log(2\pi) - \frac{T}{2}log \mid \Omega \mid -\frac{1}{2}u^T\Omega^{-1}u \tag{4.17}$$

4. Maximization of the likelihood is done by an iterative procedure. We will study the numerical optimization problem (usually some gradient method) in some detail in Chapter 4. The basic iterative procedure is in terms of the parameter updates is:

$$\Psi^{i+1} = \Psi^i + \underbrace{\left[I^i(\Psi^i)\right]^{-1}\{\frac{\partial logL}{\partial \Psi}\}} \tag{4.18}$$

where all the right hand-side of equation (4.18) is evaluated at Ψ^i. Note that $I^i(\Psi)$ is the information matrix and $\frac{\partial logL}{\partial \mathbf{\Psi}}$ is the *gradient (score)* also evaluated at $\mathbf{\Psi}^i$

5. Sometimes minus the Hessian (the second derivative of the log likelihood with no restrictions imposed; the plim of this is the information matrix) is used.

6. Most optimization schemes are based on this or some similar algorithm and iteration proceeds until some form of convergence (which may be stated in terms of a number of things).

7. To start the algorithm we need to supply initial estimates have default values and the programmer can change these.

8. As we shall see, many efficient optimization routines have variable step procedures.

4.3 Numerical Optimization Methods

- Much of the material for this Chapter is taken from Judge, Griffith, Hill, Lutkepohl and Lee. We will discuss the problem of minimization (such as minimizing a sum of squared residuals).

- However, it should be clear that maximizing the likelihood function itself can be turned into a minimizing problem just by multiplying the likelihood function by minus 1.

- Most optimization problems are **iterative** and proceed until some stopping criterion is met (this could be the number of iterations if convergence is not achieved).

Let $\boldsymbol{\theta}$ be a vector of parameters and the sequence of n iterations be indexed by $\boldsymbol{\theta}_n$, where $\boldsymbol{\theta}_1$ is the initial vector (usually this needs to be supplied and is referred to as **starting values**). At each iteration we add a vector $\boldsymbol{\zeta}_n$ –called a **step** to get $\boldsymbol{\theta}_{n+1}$

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \boldsymbol{\zeta}_n \tag{4.19}$$

Let $H(\boldsymbol{\theta})$ be the function we are trying to minimize (say a residual sum of squares). At each iteration we require $H(\boldsymbol{\theta}_n) > H(\boldsymbol{\theta}_{n+1})$.

Ideally we would like to find the $\boldsymbol{\theta}_n$ such that no improvement (reduction) in H is possible. This is not feasible from a practical point of view and so we adopt stopping rules such as (for $\epsilon > 0$, where ϵ is usually set by the researcher).

1. $(\boldsymbol{\theta}_{n+l} - \boldsymbol{\theta}_n)^T(\boldsymbol{\theta}_{n+l} - \boldsymbol{\theta}_n) < \epsilon$ inner product of estimates (STATA: 10^{-6} for coefficients)

2. $H(\boldsymbol{\theta}_n) - H(\boldsymbol{\theta}_{n+l}) < \epsilon$ for a positive l (STATA: 10^{-7})

3. $\left[\frac{\partial H}{\partial \boldsymbol{\theta}} \Big| \boldsymbol{\theta}_n \right]^T$ (STATA: can be set)

4. $\left[\frac{\partial H}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_n}\right]^T \left[\frac{\partial H}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_n}\right] < \epsilon$ inner product of gradient (STATA: 10^{-5})

5. A prescribed upper bound for the number of iterations is attained (STATA: 16,000)

6. A prescribed upper bound for the computation time is reached (old since CPU is computer dependent)

- Often we may impose one or more of these stopping rules depending upon the circumstance. For example if the minimum is not unique (1) does not guarantee termination, or if the minimum does not exist (2) and (3) might not occur so that some guaranteed stopping rule should be supplied.

- The most common optimization methods are the **gradient methods**, which we will discuss. Others like the SIMPLEX algorithm, do not use gradients and are sometimes preferred for ill-conditioned optimization problems (STATA has a simplex method but this is not available for estimating ARIMA/ARCH models).

- STATA shows some flexibility in choosing stopping rules as well as algorithms (all gradient methods) procedures for minimizing H.

  - Newton-Raphson is default
  - BHHH
  - BFGS
  - DFP

- For all problems you should always start the procedure at different starting values to determine whether the same final values are obtained.

- Often algorithms get 'stuck' at certain ranges of the parameter space.

4.4 Gradient Methods

Given a vector $\boldsymbol{\theta}_n$, we want a direction $\boldsymbol{\delta}$ (a vector) and a step length t. The idea is once we know a direction that will lower the value of the objective function, we need to decide how far to go. If we go too far, we may proceed past the trough and up to another peak; on the other hand, too short a step length is inefficient and the algorithm is slow to converge. We may illustrate the problem:

$$H(\boldsymbol{\theta}_n + t\boldsymbol{\delta}) < H(\boldsymbol{\theta}_n) \tag{4.20}$$

if $\boldsymbol{\delta}$ is 'down' $\Rightarrow$ a 'small' step in that direction will always lower the value of H. Hence we are searching for a $\boldsymbol{\delta}$ for small t such that $H(\boldsymbol{\theta}_n + t\boldsymbol{\delta})$ is a decreasing function of t:

$$\frac{d\left[H(\boldsymbol{\theta}_n + t\boldsymbol{\delta})\right]}{dt}\mid_{t=0} = \left[\frac{\partial H}{\partial \boldsymbol{\theta}}\mid \boldsymbol{\theta}_n\right]^T \left[\frac{d(\boldsymbol{\theta}_n + t\boldsymbol{\delta})}{dt}\mid_{t=0}\right] = \left[\frac{\partial H}{\partial \boldsymbol{\theta}}\mid \boldsymbol{\theta}_n\right]^T \boldsymbol{\delta}$$

which has to be less than zero (*i.e.* if $\boldsymbol{\delta}$ is 'down' then the gradient is positive and we want to lower the value of $\boldsymbol{\theta}$). Let the gradient be equal to $\boldsymbol{\gamma}_n$

$$\left[\frac{\partial H}{\partial \boldsymbol{\theta}}\mid \boldsymbol{\theta}_n\right] = \boldsymbol{\gamma}_n \tag{4.21}$$

Therefore we can choose:

$$\boldsymbol{\delta} = -\boldsymbol{P}_n\boldsymbol{\gamma}_n \tag{4.22}$$

where $\boldsymbol{P}_n$ is any **positive definite** matrix ($\boldsymbol{\gamma}_n^T\boldsymbol{P}_n\boldsymbol{\gamma}_n > 0$) for all $\boldsymbol{\gamma}_n \neq 0$. The choice of $\boldsymbol{P}_n$ is then what the various algorithms are all about.

From (4.22) we have $\boldsymbol{\gamma}_n^T\boldsymbol{\delta} = -\boldsymbol{\gamma}_n^T\boldsymbol{P}_n\boldsymbol{\gamma}_n < 0$ if $\boldsymbol{\gamma}_n \neq 0$ and if $\boldsymbol{\gamma}_n = 0$ then we have reached the minimum. The iteration equation is:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - t_n\boldsymbol{P}_n\boldsymbol{\gamma}_n \tag{4.23}$$

where t_n is the step length at iteration n.

There are many choices for $\boldsymbol{P}_n$ and no "optimal" choice over all optimization problems exists. Some algorithms choose t_n and $\boldsymbol{P}_n$ together, while others simply find any t_n that satisfies (4.20) for a given $\boldsymbol{P}_n$. Finding an optimal step length at each iteration will reduce the number of iterations but increase the computational cost for each iteration. So we shall leave the discussion of the choice of t_n to the reader (see references in Judge et al).

4.4.1 Method of Steepest Decent

$$\boldsymbol{P}_n = \boldsymbol{I}_k \tag{4.24}$$

for all iterations. In general this has slow convergence properties if the minimum is a long and narrow valley (an ill-conditioned objective function). There is no flexibility in the **direction choice** over the iterations.

4.4.2 Newton-Raphson

$$\boldsymbol{P}_n = \left[\frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\mid \boldsymbol{\theta}_n\right]^{-1} \tag{4.25}$$

Hence $\boldsymbol{P}_n$ at each iteration is set equal to the **inverse of the Hessian** which we will denote by $\boldsymbol{\mathcal{H}}_n$. To motivate this, consider a second-order Taylor series expansion of $H(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_n$:

$$H(\boldsymbol{\theta}) \cong H(\boldsymbol{\theta}_n) + \boldsymbol{\gamma}_n^T(\boldsymbol{\theta} - \boldsymbol{\theta}_n) + \ 1/2\ (\boldsymbol{\theta} - \boldsymbol{\theta}_n)^T\boldsymbol{\mathcal{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \tag{4.26}$$

The first order conditions for minimizing w.r.t $(\boldsymbol{\theta} - \boldsymbol{\theta}_n)$ imply :

$$\boldsymbol{\gamma}_n + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta} - \boldsymbol{\theta}_n) = 0 \tag{4.27}$$

which leads to:

$$\boldsymbol{\theta} = \boldsymbol{\theta}_n - \boldsymbol{\mathcal{H}}_n^{-1}\boldsymbol{\gamma}_n \tag{4.28}$$

Notes
1.If the $H(\boldsymbol{\theta})$ function is quadratic (as in a linear least squares problem) we reach the minimum in one step of length one.
2. If $\boldsymbol{\mathcal{H}}_n$ is not positive definite (outside some small neighborhood), then we have an unacceptable step which implies an **algorithm breakdown**.
3. For this algorithm, we need both analytical first and second derivatives. Other methods rely on first derivatives only. We can use numerical (less efficient) derivatives.

4.4.3 Quasi- Newton Methods

This method **does not need second derivatives** but approximates the inverse of the Hessian and adds on a correction matrix at each iteration

$$\boldsymbol{P}_{n+1} = \boldsymbol{P}_n + \boldsymbol{M}_n \tag{4.29}$$

where $\boldsymbol{M}_n$ is the correction matrix and $\boldsymbol{P}_n$ is an **approximation** to $\boldsymbol{\mathcal{H}}_n^{-1}$ at step n. In the $n+1$ iteration, $\boldsymbol{P}_{n+1}$ is the direction matrix. To start up the procedure we choose some symmetric matrix $\boldsymbol{P}_1$ and proceed inductively for reasonable choices of $\boldsymbol{P}_{n+1}$. Take a first order approximation of the gradient:

$$\boldsymbol{\gamma}_n \approx \boldsymbol{\gamma}_{n+1} + \boldsymbol{\mathcal{H}}_{n+1}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}) \tag{4.30}$$

which yields:

$$\boldsymbol{\mathcal{H}}_{n+1}^{-1}(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n) \approx (\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n) \tag{4.31}$$

so long as the Hessian is not singular. Substituting into (4.29) for $\boldsymbol{\mathcal{H}}_{n+1}^{-1} \cong \boldsymbol{P}_{n+1} = \boldsymbol{P}_n + \boldsymbol{M}_n$ gives:

$$\boldsymbol{M}_n(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n) = \boldsymbol{\eta}_n \tag{4.32}$$

$$\text{where } \boldsymbol{\eta}_n = (\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n) - \boldsymbol{P}_n(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n).$$

Notes
1. $\boldsymbol{M}_n$ is required to be symmetric so that $\boldsymbol{P}_n$ is symmetric.
2. With the number of parameters $K > 1 \Rightarrow K(K+1)/2$ different elements of $\boldsymbol{M}_n$ and therefore cannot be uniquely determined from the K equations of (4.32).
3. Different calculations of $\boldsymbol{M}_n$ lead to different algorithms.
4. $\boldsymbol{P}_1$ is commonly chosen as the identity matrix.

Rank One Correction

$$\boldsymbol{M}_n = \frac{\boldsymbol{\eta}_n \boldsymbol{\eta}_n^T}{\boldsymbol{\eta}_n^T(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)} \tag{4.33}$$

There is only one symmetric matrix of rank one that meets (4.32). Unfortunately there is no guarantee that $\boldsymbol{P}_n + \boldsymbol{M}_n$ is positive definite.

Davidon Fletcher and Powell

$$\boldsymbol{M}_n = \frac{\boldsymbol{\xi}_n \boldsymbol{\xi}_n^T}{\boldsymbol{\xi}_n^T(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)} - \frac{\boldsymbol{P}_n(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)^T \boldsymbol{P}_n}{(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)^T \boldsymbol{P}_n (\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n)} \tag{4.34}$$

$$\text{where } \boldsymbol{\xi}_n = -t_n \boldsymbol{P}_n \boldsymbol{\gamma}_n$$

Notes
1. If t_n is selected to minimize $H(\boldsymbol{\theta}_n + \boldsymbol{\xi}_n)$ for a given $\boldsymbol{\theta}_n, \boldsymbol{P}_n$ and $\boldsymbol{\gamma}_n$ then $\boldsymbol{P}_{n+1} = \boldsymbol{P}_n + \boldsymbol{M}_n$ is always positive definite.
2. There is a great deal of discussion of how to choose t_n (the step length) in the direction of $-\boldsymbol{P}_n \boldsymbol{\gamma}_n$ (see Judge et al for the references).
3. Often the **variance covariance matrix** is constructed from the **inverse Hessian**.

4.4.4 Gauss-Newton Method

This is based on another approximation for optimization problems that can be formulated as minimizing a 'sum-of-squares' type problem. Consider the regression equation:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\theta}^*) + \boldsymbol{e} = \boldsymbol{f}(\boldsymbol{\theta}^*) + \boldsymbol{e} \tag{4.35}$$

where $\boldsymbol{y}$ is a $(T \times 1)$ vector of endogenous variables, $\boldsymbol{X}$ is a $(T \times K)$ matrix of exogenous variables and $\boldsymbol{f}(\boldsymbol{\theta})$ is a $(T \times 1)$ functional vector which is presumed to be continuous and at least twice differentiable, $\boldsymbol{e}$ is a $(T \times 1)$ vector of 'residuals'. The objective function is:

$$H(\boldsymbol{\theta}) = [\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})]^T [\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta})] = \boldsymbol{e}(\boldsymbol{\theta})^T \boldsymbol{e}(\boldsymbol{\theta}) \tag{4.36}$$

The Hessian $\boldsymbol{\mathcal{H}}$ is:

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = 2\boldsymbol{Z}(\boldsymbol{\theta})^T \boldsymbol{Z}(\boldsymbol{\theta}) - 2\sum [y_t - f_t(\boldsymbol{\theta})] \left[\frac{\partial f_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big| \boldsymbol{\theta} \right] \tag{4.37}$$

where $\boldsymbol{Z}(\boldsymbol{\theta}) = \left[\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\theta}^T} \right]$ is evaluated at $\boldsymbol{\theta}$.
Since the mean of $e_t = y_t - f_t(\boldsymbol{\theta}^*)$ is assumed to be zero the second term in (4.37) is zero and the first term is assumed to be the approximation to the Hessian of $H(\boldsymbol{\theta})$. Therefore

$$\boldsymbol{P}_n = [\, 2\boldsymbol{Z}(\boldsymbol{\theta}_n)^T \boldsymbol{Z}(\boldsymbol{\theta}_n)\,]^{-1} \tag{4.38}$$

with $\boldsymbol{\gamma}_n = -2\boldsymbol{Z}(\boldsymbol{\theta}_n)^T [\, \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}_n) \,]$. Setting the step length to 1 we have:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + [\, \boldsymbol{Z}(\boldsymbol{\theta}_n)^T \boldsymbol{Z}(\boldsymbol{\theta}_n) \,]^{-1} \boldsymbol{Z}^T(\boldsymbol{\theta}_n) [\, \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}_n) \,] \tag{4.39}$$

which may be written as

$$\boldsymbol{\theta}_{n+1} = [\, \boldsymbol{Z}(\boldsymbol{\theta}_n)^T \boldsymbol{Z}(\boldsymbol{\theta}_n) \,]^{-1} \boldsymbol{Z}(\boldsymbol{\theta}_n)^T [\, \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}_n) + \boldsymbol{Z}(\boldsymbol{\theta}_n)\boldsymbol{\theta}_n \,] \tag{4.40}$$

which is just the least squares estimator for the model:

$$\bar{\boldsymbol{y}}(\boldsymbol{\theta}_n) = \boldsymbol{Z}(\boldsymbol{\theta}_n)\boldsymbol{\theta} + \boldsymbol{e} \tag{4.41}$$

where $\bar{\boldsymbol{y}}(\boldsymbol{\theta}_n) = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\theta}_n) + \boldsymbol{Z}(\boldsymbol{\theta}_n)\boldsymbol{\theta}_n$. We can see that the Gauss algorithm is just a sequence of linear regressions in which at each iteration we compute the least squares estimator of the **linear approximation** to the nonlinear model.
Notes

1. We can rewrite the algorithm in terms of something that looks more like 'minimizing a sum of squared residuals' (see Judge et al p. 961).

2. **Method of scoring** follows from the Gauss algorithm, where we set:

$$\boldsymbol{P}_n = -\left[E \frac{\partial^2 lnL}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]^{-1} \text{ evaluated at } \boldsymbol{\theta}_n.$$

   and L is the likelihood function.

3. Using the negative of the log of the likelihood function as the objective function and the Hessian is approximated by its expected value. Berndt, Hall and Hausman (BHHH) have used (STATA) :

$$\boldsymbol{P}_n = [\, \boldsymbol{Z}(\boldsymbol{\theta}_n)^T \boldsymbol{Z}(\boldsymbol{\theta}_n) \,]^{-1}$$

which does not require second derivatives.

4.5 Other Comments on Optimization

1. There are several other ways of optimizing some of which are discussed in Judge et al. One important consideration in any optimization problem is whether you have indeed achieved a minimum (or maximum). There are no hard and fast rules in this regard, but some sensitivity analysis with other starting values or other algorithms is always advisable. Certain algorithms that work well in one situation do not always do well in another. Particularly if there are long 'ridges' in the function, convergence may be tough.

2. Most programs can switch around algorithms automatically which is convenient (STATA does this)

4.5.1 Derivatives

1. Analytical
   (a) i. Many programs can take a given function and calculate the analytical derivatives directly.
   ii. Considerable computational time can often be saved with analytical derivatives.
   iii. Notice some of the algorithms above do not need second derivatives. This can have its own advantages and disadvantages too.
   iv. Not many oppportunieties in Stata to use analytical derivatives
2. Numerical Derivatives
   (a) i. Suppose we wish to evaluate the gradient (two-sided) numerically:
   $$\frac{\partial H}{\partial \theta_i} \cong \frac{[H(\theta_1, \ldots, \theta_{i-1}, \theta_i + \Delta\theta_i, \theta_{i+1}, \ldots, \theta_k) - H(\theta_1, \ldots, \theta_{i-1}, \theta_i - \Delta\theta_i, \theta_{i+1}, \ldots, \theta_k)]}{2\Delta\theta_i}$$
   if $\Delta\theta_i$ is sufficiently small.
   ii. Obviously if k is large and we are relying on numerical derivatives for a complex problem then this is going to be time computationally expensive.

4.6 Estimation of $ARMA$ Models

- STATA has long been estimating MLE models and the estimation technique is based on **Kalman filtering techniques**
- In this set-up the full likelihood (all T observations are used) and so it is worthwhile to begin estimation with a discussion of Kalman filtering
- With regard to estimating the paramaters of time series models it is computationally considerably easier if the model is entirely autoregressive.
  - In this case, the approximate (we are conditioning on the initial observations whose length depends upon the order of the AR process) MLE estimates are identical to the OLS ones. (*i.e.* maximizing the likelihood is identical to minimizing an unweighted sum of squared residuals).
- Once we introduce the MA components,
  - the model is nonlinear and the computations become increasingly burdensome
  - can see for particular examples how the optimizing is slower
- Although approaching the problem from maximizing a likelihood is somewhat restrictive (especially in light of the observation above), it nevertheless provides a *unifying* framework for the more general ARMA estimation to follow.

4.6.1 Quasi-likelihood

- Since normal support is not something we feel strongly about (Box-Jenkins has a considerable discussion on white-noise inputs that get translated into observable dependent processes), we can view the likelihood as a **quasi-likelihood**.
  - This is simply a function when maximized with respect to the parameters of interest using a specified distribution (say normal) the estimates produced have some desirable properties (much like minimizing sum of squares).
- While we make explicit our normality assumption to keep everything simple, it should be kept in mind that it is often possible (even desirable) to relax such assumptions.

4.7 Kalman Filtering: An Introduction

- The application of state space representations and the Kalman filter originally comes from the engineering literature.
- Economic applications include finance, learning behaviour, forecasting and expectation formation. In this section we shall give a simple overview of the topic.
- In Kalman filtering there are two basic equations:

$$\mathbf{y}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \quad \textbf{Measurement Equation} \tag{4.42}$$

  and the

$$\boldsymbol{\alpha}_t = T_t\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \quad \textbf{Transition Equation} \tag{4.43}$$

  where the variables have the following characteristics:
  - $\mathbf{y}_t$ is a $(N \times 1)$ vector of **observed variables**
  - $\mathbf{Z}_t$ and $\mathbf{T}_t$ (usually **time invariant**) are **fixed matrices** $(N \times M)$
  - $\boldsymbol{\alpha}_t$ is an $(m \times 1)$ **unobservable state vector**
- By assumption $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are mutually and serially independent error terms.
- We assume that $\boldsymbol{\varepsilon}_t \sim \text{NID}(\mathbf{0}, \mathbf{H}_t)$, and $\boldsymbol{\eta}_t \sim \text{NID}(\mathbf{0}, \mathbf{Q}_t)$ where $\mathbf{H}_t$ and $\mathbf{Q}_t$ may be time varying.
- Variables $\mathbf{y}_t$ are observed as a linear function of the unobservable $\boldsymbol{\alpha}_t$ and the error term $\boldsymbol{\varepsilon}_t$ in (4.42).
- The transition equation is always in Markov, **state space** or **companion form** (with one lag).

- We shall see that any ARMA model may be rewritten in the state space form (4.43).
- The transition equation characterizes the evolution of $\boldsymbol{\alpha}_t$ over time with a starting value $\boldsymbol{\alpha}_0$ assumed to be independent of either the $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$.
- Notice, if we have some initial conditions, then (4.42) and (4.43) completely specify the model.

4.8 Goal of Kalman Filtering (Predicting and Updating)

In this exercise we wish to estimate (or learn) $\boldsymbol{\alpha}_t$, which we cannot observe directly by combining optimally:

1. The estimated prediction of $\boldsymbol{\alpha}_t$, say $\mathbf{a}_{t/t-1}$, the forecast based on information at time $t-1$.
2. The current observation on $\mathbf{y}_t$ (constitutes the new information).

- We note that since $\boldsymbol{\alpha}_t$ is assumed to be stochastic that we cannot speak about the properties of estimators in the same way as the fixed parameter world.
- Instead, the language is more similar to the prediction or forecasting terminology (i.e. minimum mean square linear estimator: MMSE).

4.9 Motivation of Kalman Filtering Problem

4.9.1 Forecasting with Measurement Error

- Reinterpret the transition equation (4.43) with $\mathbf{X}$ serving as $\boldsymbol{\alpha}$.
- Hence
$$\mathbf{X}_t = \mathbf{T}_t\mathbf{X}_{t-1} + \boldsymbol{\eta}_t \tag{4.44}$$
which is a simple autoregressive model.
- Now the problem arises; if $\mathbf{X}_t$ is assumed to be unobservable and hence we use the observation equation:
$$\mathbf{y}_t = \mathbf{Z}\mathbf{X}_t + \boldsymbol{\varepsilon}_t \tag{4.45}$$
where $\mathbf{Z}$ is a **known parameter** (fixed in t here) and $\mathbf{y}_t$ is observed.
- This is a signal extraction problem: we wish to learn about $\mathbf{X}_t$ from:
  - the current $\mathbf{y}_t$ (which has two components both of which are unobservable)
  - the forecast of $\mathbf{X}_{t/t-1}$ from the filter prediction.

4.9.2 Time Varying Parameters

- Reinterpret the measurement equation (4.42) as the more familiar regression notation with the stochastic $\boldsymbol{\alpha}$ as $\boldsymbol{\beta}$:

$$\mathbf{y}_t = \mathbf{Z}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t \tag{4.46}$$

  with a transition equation for $\boldsymbol{\beta}_t$ as:

$$\boldsymbol{\beta}_t = \mathbf{T}_t\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t. \tag{4.47}$$

- This is often the interpretation used to motivate the Kalman approach as is the kind of Kalman filtering problem that STATA is capable of handling.

4.10 State Space Representation of an ARMA Model

- The general ARMA(p, q) structure:

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q} \tag{4.48}$$

  can be written in state space form (4.43) as:

$$\boldsymbol{\alpha}_t = \begin{pmatrix} \phi_1 & 1 & 0 & \ldots & 0 \\ \phi_2 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \\ \phi_{m-1} & 0 & 0 & \ldots & 1 \\ \phi_m & 0 & 0 & \ldots & 0 \end{pmatrix} \boldsymbol{\alpha}_{t-1} + \begin{pmatrix} 1 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_{m-1} \end{pmatrix} \boldsymbol{\varepsilon}_t \tag{4.49}$$

  where $m = \max(p, q+1)$ and therefore $\phi_i = 0$ for $i > p$ and $\theta_i = 0$ for $i > q$.

- This may be regarded as the transition equation (4.43) in which $\mathbf{T}_t$ is a constant $(m \times m)$ matrix and $\mathbf{Q}_t = \sigma^2 \; \forall \; t$.

- The first element in $\boldsymbol{\alpha}_t$ will be identically equal to $\mathbf{y}_t$.

- To appreciate that nothing has been lost or added in (4.49) let us deal with a couple of special cases.

4.10.1 $MA(1)$

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} \tag{4.50}$$

- The Markovian representation of the $MA(1)$ process:

$$\boldsymbol{\alpha}_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \boldsymbol{\alpha}_{t-1} + \begin{pmatrix} 1 \\ \theta \end{pmatrix} \varepsilon_t \tag{4.51}$$

where $\boldsymbol{\alpha}_t = \begin{pmatrix} \alpha_{1t} & \alpha_{2t} \end{pmatrix}^T$ and $\alpha_{2t} = \theta\varepsilon_t$ and $\alpha_{1t} = \alpha_{2t-1} + \varepsilon_t = \varepsilon_t + \theta\varepsilon_{t-1}$ and the measurement equation needs to extract the first element of the state vector (namely ε_t):

$$y_t = \mathbf{z}_t^T \boldsymbol{\alpha}_t \quad t = 1, \ldots, T \tag{4.52}$$

where $\mathbf{z}_t^T = (1, 0)$.

- Notice that there is no error in the measurement equation.

4.10.2 *ARMA*(2, 1)

- $$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t + \theta\varepsilon_{t-1}. \tag{4.53}$$

- This can be written as:

$$\begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta\varepsilon_t \end{pmatrix} = \begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix} \begin{pmatrix} y_{t-1} \\ \phi_2 y_{t-2} + \theta\varepsilon_{t-1} \end{pmatrix} + \begin{pmatrix} 1 \\ \theta \end{pmatrix} \varepsilon_t \tag{4.54}$$

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta\varepsilon_t \end{pmatrix} \tag{4.55}$$

  which is the same form as (4.42) ($y_t = \mathbf{z}_t^T \boldsymbol{\alpha}_t + \varepsilon_t$)

- with $\boldsymbol{\alpha}_t = \begin{pmatrix} y_t \\ \phi_2 y_{t-1} + \theta\varepsilon_t \end{pmatrix}, \mathbf{T}_t = \mathbf{T} = \begin{pmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{pmatrix}$ and $\mathbf{z}_t^T = (1, 0)$ again with no measurement error in (4.42), $\varepsilon_t = 0$.

- STATA uses a slightly different set up for its Kalman filtering problem based on Hamilton (1994).

- As an exercise do an ARMA(1,1).

4.11 Minimum Mean Square Error Estimation

- The **Kalman filter** is a tool to estimate the state vector $\boldsymbol{\alpha}_t$ in an optimal way and to update the estimate when an new observation ($\mathbf{y}_t$) becomes available.

4.11.1 *Digression on Generalized Least Squares*

- Notice the error term has a variance covariance that is not spherical (not the identity matrix).

- In this case, our standard theory says that we should use some generalized least squares (GLS) procedure to estimate the parameters of the model efficiently.

- Suppose:
$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{4.56}$$
where $\mathbf{y}$ is a $(T \times 1)$ matrix, $\mathbf{Z}$ is a set of fixed regressors $(T \times m)$, and $\boldsymbol{\varepsilon}$ is a $(T \times 1)$ vector of disturbances with mean $\mathbf{0}$ and variance covariance matrix $\sigma^2\boldsymbol{\Omega}$.

- Assume that $\boldsymbol{\Omega}$ is positive definite and **known**. If $\boldsymbol{\alpha}$ were fixed (as in the general classical linear regression model), then feasible Aitken estimation would be BLUE:
$$\tilde{a} \;=\; (\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{y} \tag{4.57}$$

- However with a **random** $\boldsymbol{\alpha}$, we must talk in terms of the **distribution of the estimation error** $(\tilde{\mathbf{a}} - \boldsymbol{\alpha})$. Point wise consistency has no interpretation.

- We write:
$$(\tilde{\mathbf{a}} - \boldsymbol{\alpha}) \sim WS\left[\mathbf{0}, \sigma^2(\mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}\right] \tag{4.58}$$
where WS (following Harvey) stands for wide sense (unconditional means and variances).

- Note both sides of $(\tilde{\mathbf{a}} - \boldsymbol{\alpha})$ have a distribution. Harvey refers to the mean zero in $(\tilde{\mathbf{a}} - \boldsymbol{\alpha})$ as **unconditionally unbiased** or (**u-unbiased**).

- We note that any other **linear** estimator of (4.56), say $\hat{\mathbf{a}}$ will have $V(\hat{\mathbf{a}}) - V(\tilde{\mathbf{a}}) \;=\; M$, a positive semi-definite matrix.

- Hence (4.57) is a minimum variance estimator among the class of all other **u**-unbiased linear estimators and we say that it is **minimum mean square linear estimator** (MMSLE).

- If we add the assumption of normality, then we are not confined to the class of linear estimators and we have MMSE - **minimum mean square estimator.**

4.12 The Kalman Filter

- Now moving back to our original problem. We will simplify the model and consider only the scalar case with $\mathbf{T}_t \;=\; \mathbf{T}$ (time invariant)
$$\mathbf{y}_t \;=\; \mathbf{z}_t^T\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \;\text{ Measurement equation } \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \sigma^2\mathbf{h}_t) \tag{4.59}$$
$$\boldsymbol{\alpha}_t \;=\; \mathbf{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t \;\text{ Transition equation } \boldsymbol{\eta}_t \sim N(\mathbf{0}, \sigma^2\mathbf{Q}) \tag{4.60}$$
where we have introduced σ^2 (which is arbitrary) to follow the notation of Harvey.

- Note the constant variance assumption in (4.60).

- We will also follow Harvey and denote $\mathbf{a}_t$ as our best guess of $\boldsymbol{\alpha}_t$ based upon all information up to and including the current observation on $\mathbf{y}_t$.
- Thus $\mathbf{a}_t$ is the MMSLE of $\boldsymbol{\alpha}_t$ at time t and $\mathbf{a}_{t/t-1}$ will be the MMSLE of $\boldsymbol{\alpha}_t$ at time $t-1$.
- The / convention tells us what we are conditioning our guess of $\boldsymbol{\alpha}$ on.

4.12.1 Prediction

- Recall we are trying to combine different pieces of information together optimally.
- It is convenient to take a Bayesian perspective on the problem.
- Suppose we have an initial prior (or guess) on $\boldsymbol{\alpha}_0$, say $\mathbf{a}_0$, such that:

$$(\mathbf{a}_{0/0} - \boldsymbol{\alpha}_0) \sim WS(\mathbf{0}, \sigma^2 \mathbf{P}_0). \tag{4.61}$$

- The filter combines the prediction $\mathbf{a}_{1/0}$ (using the transition equations) with the likelihood on the first observation on $\mathbf{y}_1$ to yield $\mathbf{a}_1$.
- This represents our best estimate of $\boldsymbol{\alpha}_1$ by combining optimally last period's best estimate of $\boldsymbol{\alpha}_1$ with the new information represented by the observation on $\mathbf{y}_1$.
- We continue to follow this strategy so that $\mathbf{a}_{2/1}$ is the prediction of $\boldsymbol{\alpha}_2$ which is then combined with the new observation $\mathbf{y}_2$. Thus $\mathbf{a}_{t/t-1}$ serves as an **evolving prior**.
- Consider how to combine $\mathbf{a}_{t/t-1}$ and $\mathbf{y}_t$ optimally.
  - If we apply the transition equation, it follows that given $\mathbf{a}_{t-1}$ (this is our best guess of what $\boldsymbol{\alpha}_{t-1}$ was using information $t-1$) that:

$$\mathbf{a}_{t/t-1} \;=\; \mathbf{T}\mathbf{a}_{t-1} \tag{4.62}$$

- The covariance of the prediction error $E\left[(\mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t)(\mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t)^T\right]$ is:

$$\mathbf{P}_{t/t-1} \;=\; \mathbf{T}\mathbf{P}_{t-1}\mathbf{T}^T + \sigma^2\mathbf{Q} \tag{4.63}$$

and so we write:

$$\mathbf{a}_{t/t-1} - \boldsymbol{\alpha}_t \sim WS(\mathbf{0}, \sigma^2 \mathbf{P}_{t/t-1}) \tag{4.64}$$

- We may relate this to the $\mathbf{y}_t$ so that the MMSLE of $\mathbf{y}_t$ is:

$$\tilde{\mathbf{y}}_{t/t-1} \;=\; \mathbf{z}_t^T \mathbf{a}_{t/t-1} \tag{4.65}$$

with a prediction error which is decomposed into two orthogonal components:

$$\boldsymbol{\nu}_t \;=\; \mathbf{y}_t - \tilde{\mathbf{y}}_{t/t-1} = \mathbf{z}_t^T(\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1}) + \boldsymbol{\varepsilon}_t \tag{4.66}$$

- Given that $E\left[\boldsymbol{\alpha}_t - \mathbf{a}_{t/t-1}\right] \;=\; E\left[\boldsymbol{\varepsilon}_t\right] \;=\; \mathbf{0}$, then $E\left[\boldsymbol{\nu}_t\right] \;=\; \mathbf{0}$.
- The orthogonality in (4.66) implies:

$$V\left[\boldsymbol{\nu}_t\right] \;=\; \sigma^2 \mathbf{z}_t^T \mathbf{P}_{t/t-1}\mathbf{z}_t + \sigma^2 \mathbf{h}_t \;=\; \sigma^2 \mathbf{f}_t \tag{4.67}$$

- Keep in mind that (4.64) and (4.66) are the **prediction errors**.

4.12.2 Updating

- We want to examine how new information in $\mathbf{y}_t$ is incorporated with the optimal predictor $\mathbf{a}_{t/t-1}$. Harvey motivates the problem in terms of a **mixed estimation** context.
- However, the following result is insightful and is stated as a lemma.

Lemma 1 *If X and Y are jointly normal random variables, then the distribution of Y* **conditional** *on X is normal with a mean*

-

$$E\left[Y|X\right] \;=\; E\left[Y\right] + V_{XY}V_{XX}^{-1}(X - E\left[X\right])$$

$$V[Y|X] \;=\; V_{YY} - V_{YX}V_{XX}^{-1}V_{XY}$$

- This is really just an example of linear projection theory.
- Now we may use the results of the lemma to develop the updating equations from (4.42) and (4.43):

$$\mathbf{a}_t \;=\; \mathbf{a}_{t/t-1} + \mathbf{P}_{t/t-1}\mathbf{Z}_t^T\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t\mathbf{a}_{t/t-1}) \tag{4.68}$$

$$\mathbf{P}_t \;=\; \mathbf{P}_{t/t-1} - \mathbf{P}_{t/t-1}\mathbf{Z}_t^T\mathbf{F}_t^{-1}\mathbf{Z}_t\mathbf{P}_{t/t-1} \tag{4.69}$$

$$\mathbf{F}_t \;=\; \mathbf{Z}_t\mathbf{P}_{t/t-1}\mathbf{Z}_t^T + \mathbf{H}_t \tag{4.70}$$

where we have used the vector notation (it just seems more elegant).

- The prediction error

$$\boldsymbol{\nu}_t \;=\; \mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t/t-1} \tag{4.71}$$

with a mean of zero and a variance $E\left[\boldsymbol{\nu}_t \boldsymbol{\nu}_t^T\right] \;=\; \mathbf{F}_t$.

- Equation (4.68) is the Kalman filter and should be compared against the first equation of the lemma.
- The first term in (4.68) is simply the optimal prediction of $\boldsymbol{\alpha}_t$ given information at time $t-1$ and the second part represents the amount that we adjust the forecast in light of the new information from $\mathbf{y}_t$.
- The prediction error $\boldsymbol{\nu}_t \;=\; \mathbf{y}_t - \mathbf{Z}_t \mathbf{a}_{t/t-1}$ appears in (4.68) and is weighted by the **Kalman gain:**

$$\mathbf{P}_{t/t-1} \mathbf{Z}_t^T \mathbf{F}_t^{-1} \;=\; \mathbf{P}_{t/t-1} \mathbf{Z}_t^T (\mathbf{Z}_t \mathbf{P}_{t/t-1} \mathbf{Z}_t^T + \mathbf{H}_t)^{-1} \tag{4.72}$$

- Two effects are apparent

1. If $\mathbf{H}_t$ is large, implying the relation between $\mathbf{y}_t$ and the $\boldsymbol{\alpha}$ is a noisy one, and therefore $\mathbf{y}_t$ provides a relatively poor signal as to what is happening to $\boldsymbol{\alpha}_t$. In this case, the Kalman gain is small.
2. If $\mathbf{P}_{t/t-1}$ is large, implying a large prediction error for $\boldsymbol{\alpha}$ and hence indicating that $\boldsymbol{\alpha}_t$ may differ substantially from $\mathbf{a}_{t/t-1}$; this suggests that we should weight the information from the current observation $\mathbf{y}_t$ a great deal. In this case the Kalman gain is large.

- Equation (4.69) is similar to (4.68) and is just the updating equation for the variance.
- We also note that in some circumstances (as in the ARMA examples) there are no errors in the measurement equation so that $\mathbf{H}_t \;=\; \mathbf{0}$.
- Note if prior information is available $\boldsymbol{\alpha}_0 \sim WS(\mathbf{a}_0, \mathbf{P}_0)$ where $\mathbf{a}_0$ and $\mathbf{P}_0$ are known then we can obtain MMSLE of $\boldsymbol{\alpha}_T$ based on all T observations.
- To get things started we need the initial conditions.

4.12.3 Specification and Estimation

- To make the Kalman filter exercise operational we need to specify

1. a suitable state space model
2. starting values $\mathbf{a}_0$ and $\mathbf{P}_0$ (supplied by STATA)

- That is, we must select some representation of the time series which involves modeling, judgment, and testing

4.13 Prediction Error Decomposition: Estimating ARMA Models and MLE

- This representation of the time series problem allows for exact maximum likelihood procedures by the Kalman Filter (owing to the usual equivalence)
- This section looks at the problem but it should be obvious if ARMA models can be written in state-space form as earlier, then all we need are initial conditions to get going (see Stata manula for the initial conditions)
- Consider a set of dependent observations on $y : y_1, \ldots, y_T$ with $y \sim N(\mu, \sigma^2 V)$.
- The log likelihood (or we may still interpret this as a joint density) is:

$$logL(y) \;=\; -(T/2)log2\pi - (T/2)\log\sigma^2 - (1/2)\log \mid V \mid -(1/2)\sigma^{-2}(y-\mu)^T V^{-1}(y-\mu) \tag{4.73}$$

which may be factored into:

$$logL(y) \;=\; logL(y_1, \ldots y_{T-1}) + \log l(y_T/y_{T-1}, y_{T-2}, \ldots, y_1) \tag{4.74}$$

- The last term is the **conditional** distribution of y_T given the observations

$$y_{T-1}, y_{T-2}, \ldots y_1.$$

  (Recall the joint probability is equal to the marginal times the conditional probability).

- The notational convention is to use L to refer to the joint likelihood and l for the conditional likelihood.
- The problem of **estimating $\mathbf{y_T}$** given that $y_{T-1}, y_{T-2}, \ldots y_1$ are known is direct.
- Following the notation of Harvey, let $\tilde{y}_{T/T-1}$ be such an estimator.
- The **prediction error** may be decomposed into two parts:

$$y_T - \tilde{y}_{T/T-1} \;=\; [y_T - E[y_T/y_{T-1}, \ldots, y_1]] + \Big[E[y_T/y_{T-1}, \ldots, y_1] - \tilde{y}_{T/T-1}\Big] \tag{4.75}$$

where $E[y_T/y_{T-1}, \ldots, y_1]$ is the mean of the conditional distribution of y_T.

(Obviously one way to actually estimate this is a least squares projection on the lagged observations).

- Using this **orthogonal** decomposition we obtain the **mean squared error** (MSE) as:

$$MSE(\tilde{y}_{T/T-1}) \;=\; Var[y_T/y_{T-1},\ldots,y_1] + E\left[\left\{\tilde{y}_{T/T-1} - E[y_T/y_{T-1},\ldots,y_1]\right\}^2\right] \tag{4.76}$$

- If we wish to minimize the MSE w.r.t. $\tilde{y}_{T/T-1}$, we note that the first term in (4.76) does not involve it, so that the **minimum mean squared estimator** ($MMSE$) of y_T conditional on $y_{T-1},\ldots,y_1$ is:

$$\tilde{y}_{T/T-1} \;=\; E\left[y_T/y_{T-1},\ldots,y_1\right] \tag{4.77}$$

  with a prediction error variance of $E\left[(y_T - \tilde{y}_{T/T-1})^2\right]$ which, given (4.77) is equal to $Var\left[y_T/y_{T-1},\ldots,y_1\right]$, denoted $\sigma^2 f_T$ (the conditional variance).

- We can write the conditional likelihood for y_T of the second term from (4.74) as:

$$\log l(y_T/y_{T-1},y_{T-2},\ldots y_1) = -(\frac{1}{2})log2\pi - (\frac{1}{2})log\sigma^2 - (\frac{1}{2})log f_T - \frac{(\frac{1}{2})\sigma^{-2}(y_T - \tilde{y}_{T/T-1})^2}{f_T} \tag{4.78}$$

  which can be interpreted as the distribution of the prediction error of $y_T - \tilde{y}_{T/T-1}$.

- We can see that this type of decomposition can be repeated w.r.t. to observation $T-1$ and then observation $T-2$ and so on.

- This leads to a likelihood:

$$\log L(y) \;=\; \sum_{t=2}^{T} logl(y_t/y_{t-1},\ldots,y_1) + \log l(y_1) \tag{4.79}$$

- This shows that the mean of y_t, conditional on the previous observations, is $\tilde{y}_{t/t-1}$ which is the $MMSE$ given the previous observations.

- Each of the **conditional distributions** is the distribution of the error associated with the **optimal** predictor.

- Letting

$$\upsilon_t \;=\; y_t - \tilde{y}_{t/t-1} \tag{4.80}$$

  we may write the likelihood into the **joint** distribution of the T independent prediction errors:

$$\log L(y) \;=\; -(T/2)log2\pi - (T/2)log\sigma^2 - (1/2)\sum_{t=1}^{T} log f_t - (1/2)\sigma^{-2}\sum_{t=1}^{T}\frac{\upsilon_t^2}{f_t} \tag{4.81}$$

4.13.1 Exact Likelihood

- Now initially we said that this kind of decomposition is useful in estimating the **exact likelihood**.

- The prediction error decomposition can be viewed in terms of a Cholesky decomposition on the V^{-1} in the likelihood at the start of this section.

- Let $\bar{L}$ be a lower diagonal matrix with ones on the main diagonal and

$$V^{-1} = \bar{L}^T D \bar{L} \tag{4.82}$$

where

$$D = \text{diag}(f_1^{-1}, \ldots, f_T^{-1}) \tag{4.83}$$

- This is a unique factorization with the prediction errors

$$v = \bar{L}y \tag{4.84}$$

- The transformation $\bar{L}y$ has a Jacobian of unity ($\mid \bar{L}y \mid = 1$) so that the joint distribution of

$$v = (v_1, v_2, \ldots v_T)^T$$

is (4.81). Also

$$\mid V^{-1} \mid = \mid \bar{L} \mid \times \mid D \mid \times \mid \bar{L} \mid = \mid D \mid \tag{4.85}$$

and therefore

$$log \mid V \mid = \sum log f_t \tag{4.86}$$

- Hence the Cholesky decomposition can be used to factor V^{-1} numerically and computing prediction errors from $\bar{L}y$ directly..

- Notice at this level of generality, the prediction error v_t does not need to have a constant conditional variance.

- However, adding this assumption, we may then start to identify the forecast error as white noise i.e. $v_t = \epsilon_t$ and hence $f_t = 1$ for all t.

- We are then left with the more familiar likelihood:

$$\text{logL}(\boldsymbol{y}) = -(T/2)log2\pi - (T/2)log\sigma^2 - (1/2)\sigma^{-2}\sum_{t=1}^{T}\epsilon_t^2 \tag{4.87}$$

- By finding such transforms, we see that maximizing (4.87) is equivalent to minimizing a sum of squared errors problem which is done using the Kalman filter!

4.14 Estimating by Non-Linear Least Squares

- An older tradition of estimating these models was to base estimation on non-linear least squares
- Explains why all of the OLS proofs asympotically go through and why MLE here is just a convenience for distribution theory (normality and such)
- We provide the background here since many packages other than STATA approach this problem for this perspective (RATS for instance)

4.14.1 The $AR(p)$ Process

- The log likelihood for the T observations is:

$$LogL(\boldsymbol{y}) = \sum_{t=p+1}^{T} log(y_t/y_{t-1}, \ldots, y_1) + logL(\boldsymbol{y}_p) \tag{4.88}$$

- The leading term is the logarithm of the joint distribution of the corresponding $\epsilon_{p+1}, \ldots, \epsilon_T$ and the second term is the joint distribution of the first p observations $\boldsymbol{y}_p = (y_1, \ldots, y_p)^T$ (often the likelihood will be conditional on these and we omit them in estimation).
- If the covariance matrix of the first p observations is $\sigma^2 \boldsymbol{V}_p$, the full log likelihood $LogL(\boldsymbol{y})$ is:

$$-(\frac{1}{2})Tlog(2\pi) - \frac{1}{2}Tlog\sigma^2 - \frac{1}{2}log \mid \boldsymbol{V}_p \mid -\frac{1}{2}\sigma^{-2}\left[\boldsymbol{y}_p^T \boldsymbol{V}_p^{-1} \boldsymbol{y}_p + \sum_{t=p+1}^{T} (y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p})^2\right]$$

- We may in the usual way concentrate out σ^2, but this is still a nonlinear problem because of the initial p parameters.
- The estimation procedure is a **conditional on the first p observations** which is just a least squares problem
- We can (as in Fuller, $p328$) make explicit assumptions for the model as in the $AR(1)$ case.
- Consider the model in T observations as

$$y_1 = \mu_1$$

$$y_t = \phi_1 y_{t-1} + \epsilon_t \qquad t = 2, \ldots, T$$

where the vector $(\mu_1, \epsilon_2, \ldots, \epsilon_T)$ is distributed as a multivariate normal with zero mean and a variance covariance matrix $\boldsymbol{\Sigma}$ (use unconditonal distribution for μ_1):

$$\mathbf{\Sigma} = \begin{bmatrix} (1-\phi_1^2)^{-1}\sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- This assumption leads to the familiar likelihood:

$$LogL = -(\frac{1}{2})Tlog(2\pi) - \frac{1}{2}Tlog\sigma^2 + \frac{1}{2}log(1-\phi_1^2) - \frac{1}{2}\sigma^{-2}\left[(1-\phi_1^2) + \sum_{t=2}^{T}(y_t - \phi_1 y_{t-1})^2\right]$$

- Hence we maximize this likelihood with respect to ϕ_1 and σ^2.
- However the most common technique is to condition on (treat as given or fixed) the initial observations. In this case, the problem is a familiar **minimizing the residual sum of squares**.

$$S(\boldsymbol{\phi}) = \sum_{t=p+1}^{T}(y_t - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p})^2 \tag{4.89}$$

- Hence conditional maximum likelihood is obtained by the OLS regression of y_t on its past values $y_{t-1}, .., y_{t-p}$.

Notes

1. In large samples it makes little practical difference if the exact likelihood is used. (Although there are some pathological examples in the literature).
2. Introducing a non-zero mean (called a drift) is simple

$$S(\boldsymbol{\phi}) = \sum (y_t - \mu - \phi_1(y_{t-1} - \mu) - \dots - \phi_p(y_{t-p} - \mu))^2 \tag{4.90}$$

3. In (4.90) is $\bar{\boldsymbol{y}}$, the sample mean, a consistent estimate of μ. (No- Why?).
4. We can use the least squares estimate of σ^2:

$$s^2 = \frac{1}{T-p}\sum \left(y_t - \hat{\mu} - \hat{\phi}_1(y_{t-1} - \hat{\mu}) - \dots - \hat{\phi}_p(y_{t-p} - \hat{\mu})\right)^2 \tag{4.91}$$

   where the $\hat{\mu}$ and $\hat{\phi}_1, \dots, \hat{\phi}_p$ are the OLS estimates. We can use the MLE estimate and divide by T.

5. In estimating by least squares we have not constrained the process to be stationary. That is, the estimated coefficients of the polynomial in the lag operator may lie outside the unit circle. In exact MLE we have constrained the problem so that nonstationary values are not admissible. To see this, notice that in the exact $AR(1)$ likelihood above as $\mid \phi_1 \mid$ approaches 1 the likelihood approaches minus infinity.

4.14.2 Estimating Moving Average Processes by Non-Linear Least Squares

- In general, moving average processes in econometric modeling tend to be of low order (Unlike AR models, the order of the MA is typically 1 or 2).
- This likely reflects the relative ease of estimating AR models over MA ones.
- As before, the estimation method is maximum likelihood and again there is conditioning so that the problem can be transferred into minimizing sums of squares.

4.14.3 Estimating a $MA(1)$ Process

The $MA(1)$ model (see Fuller, p.344 or Harvey 124-125) :

$$y_t \;=\; \epsilon_t + \theta\epsilon_{t-1} \tag{4.92}$$

We may rewrite (or in the case of Harvey reinterpret) (4.92):

$$\epsilon_t \;=\; -\theta\epsilon_{t-1} + y_t \tag{4.93}$$

The ϵ_t will eventually be treated as a **residual** term, whose value depends on the choice of θ (and not a disturbance term).

Using some recursive results obtained earlier we may write (4.93) as:

$$\epsilon_t = \sum_{j=0}^{\infty}(-\theta)^j y_{t-j} \tag{4.94}$$

which for estimation purposes can be expressed as:

$$y_t \;=\; -\sum_{j=1}^{\infty}(-\theta)^j y_{t-j} + \epsilon_t \tag{4.95}$$

Clearly this is a very nonlinear problem (not to mention the infinite lag).
From (4.94) we see that ϵ_t is a function of the θ and the data $y_t, y_{t-1}, \ldots$ which may be made explicit :

$$\epsilon_t(\boldsymbol{Y};\theta) = \sum_{j=0}^{\infty}(-\theta)^j y_{t-j} = \sum_{j=0}^{t-1}(-\theta)^j y_{t-j} + (-\theta)^t\epsilon_0 \tag{4.96}$$

where $\epsilon_0 = \sum_{j=0}^{\infty}(-\theta)^j y_{-j}$.

Hence, coming full circle we denote the $MA(1)$ process as:

$$y_t \;=\; \theta\epsilon_{t-1}(\boldsymbol{Y};\theta) + \epsilon_t \tag{4.97}$$

Note
1. For each value of θ we **build** up a set of residuals given by (4.96)

2. A typical strategy is to set $\epsilon_0 = 0$ (more correctly the estimator of ϵ_0 is zero –which is a unbiased estimate of it)
3. We can use a Gauss-Newton type algorithm (Fuller, $p343$):

We first need a starting estimate of θ. For instance we can use the estimated autocorrelation coefficient $\hat{\rho}(1)$ and the following rule:

$$\begin{aligned} \hat{\theta} &= [2\hat{\rho}(1)]^{-1}\left\{1 - [1 - 4\hat{\rho}^2(1)]^{\frac{1}{2}}\right\} && 0 <\mid \hat{\rho}(1) \mid \leq 0.5 \\ &= -1 && \hat{\rho}(1) < 0.5 \\ &= 1 && \hat{\rho}(1) > 0.5 \\ &= 0 && \hat{\rho}(1) = 0 \end{aligned}$$

We note that if $\hat{\rho}(1)$ lies outside the range of 0.5 by a significant amount then the $MA(1)$ model is suspect.

From this we have a well-defined nonlinear problem. Following the notation earlier for the Gauss-Newton method, we set $Z_t(\theta) = \partial\epsilon_t(\mathbf{Y};\theta)/\partial\theta$ evaluated at the current estimate of θ. Therefore, we may write the iterative formula for θ_{n+1} as:

$$\theta_{n+1} = \theta_n + \frac{\sum_1^T Z_t(\theta_n)\epsilon_t(\mathbf{Y},\theta_n)}{\sum [Z_t(\theta_n)]^2} \tag{4.98}$$

Notice that (4.98) is just a specialization of (4.39). The gradient also satisfies the difference equation:

$$Z_t(\theta_n) = \epsilon_{t-1}(\mathbf{Y},\theta_n) - \theta_n Z_{t-1}(\theta_n) \tag{4.99}$$

which may aid in the calculation to build up the gradient. This can be seen by differentiating (4.93).

4.15 The MA(q) Process and NLS Estimation of ARMA(p,q) Processes

Basically the same sorts of rules apply for the higher order MA processes. We first write the model as ($t = 1, \ldots T$):

$$\epsilon_t = -\theta_1\epsilon_{t-1} - \theta_2\epsilon_{t-2} - \ldots - \theta_{t-q}\epsilon_{t-q} + y_t \tag{4.100}$$

and then set all the initial (q) estimates of $\epsilon_{1-q} = \epsilon_{2-q} = \ldots = \epsilon_0 = 0$.
From an initial value for the $\theta's$ a set of conditional residuals is built up in the same way (see Fuller, $p349-351$) and a nonlinear optimization routine like Gauss-Newton is applied.

Also the addition of the AR portion, the ARMA(p,q), possess no additional problems:

$$\epsilon_t = y_t - \phi_1 y_{t-1} - \ldots - \phi_p y_{t-p} - \theta_1\epsilon_{t-1} - \ldots - \theta_q\epsilon_{t-q} \tag{4.101}$$

$$t = p+1, \ldots, T \text{ and } \epsilon_p = \epsilon_{p-1} = \ldots = \epsilon_{p-q+1} = 0$$

This will in general involve $p+q$ derivatives. In Box and Jenkins (p.237) there are 2 recursion formulas for building up the appropriate regressors for the minimization problem for such algorithms as Gauss-Newton.

4.16 Stationarity and Invertibility

- As mentioned earlier, if the exact likelihood is used for the estimation then stationarity of the process is guaranteed for $AR(1)$ processes.
- However, for higher order processes or the approximate procedures outlined, the stationarity conditions may indeed be violated.
- In general it is quite difficult to impose the stationarity conditions directly in estimation and are typically checked only after estimation.
- Notice that the stationarity conditions involve **inequality restrictions** and do not fit into standard hypothesis testing procedures.
- Invertibility is not necessarily imposed. Recall that non-invertibility does not imply that the process is nonstationary. Invertibility is really a problem of identification.
- For every noninvertible representation, there is an invertible representation obtained by taking the roots inside the unit circle (one that has the same autocorrelation function).