Parameter Estimation

Leonid Kogan

MIT, Sloan

15.450, Fall 2010

Outline

Outline

Statistics Review: Parameter Estimation

- Sample of observations $X = (x_1, ..., x_T)$ with joint distribution $p(X, \theta_0)$.
- Estimator $\widehat{\theta}$ is a function of the sample: $\widehat{\theta}(X)$.
- Estimator is *consistent* if

$$\text{plim}_{T\to\infty}\,\widehat{\theta} = \theta_0$$

- Estimator is *unbiased* if

$$\mathrm{E}[\widehat{\theta}] = \theta_0$$

- An α confidence interval for the i'th coordinate of the parameter vector, $\theta_{0,i}$, is a stochastic interval

$$(\widehat{\theta}_i^L, \widehat{\theta}_i^R) \;\text{ such that }\; \text{Prob}\left[(\widehat{\theta}_i^L, \widehat{\theta}_i^R) \text{ covers } \theta_{0,i}\right] = \alpha$$

Probability Review: LLN and CLT

- Law of Large Numbers (LLN) states that if x_t are IID random variables and $\mathrm{E}[x_t] = \mu$, then

$$\mathrm{plim}_{T\to\infty} \frac{\sum_{t=1}^{T} x_t}{T} = \mu$$

- plim is limit in probability. $\mathrm{plim}_{n\to\infty} x_n = y$ means that for any $\delta > 0$, $\mathrm{Prob}[|x_n - y| > \delta] \to 0$.
- Central Limit Theorem (CLT) states that if x_t are IID random vectors with mean vector μ and var-cov matrix Ω, then

$$\frac{\sum_{t=1}^{T} (x_t - \mu)}{\sqrt{T}} \Rightarrow \mathcal{N}(0, \Omega)$$

- "$\Rightarrow$" denotes convergence in distribution. $x_n \Rightarrow y$ means that the corresponding cumulative distribution functions $F_{x_n}(\cdot)$ and $F_y(\cdot)$ have the property

$$F_{x_n}(z) \to F_y(z) \quad \forall z \in \mathbb{R}, \text{ s.t., } F_y \text{ is continuous at } z$$

Example

- We observe a sample of IID observations x_t, $t = 1, \ldots, T$ from a Normal distribution $\mathcal{N}(\mu, 1)$.
- We want to estimate the mean μ.
- A commonly used estimator is the sample mean:

$$\widehat{\mu} = \widehat{\mathrm{E}}[x_t] \equiv \frac{1}{T} \sum_{t=1}^{T} x_t$$

- This estimator is consistent by the LLN: $\mathrm{plim}_{T \to \infty}\, \widehat{\mu} = \mu$.
- How do we derive consistent estimators in more complex situations?

Approaches to Estimation

- If probability law $p(X, \theta_0)$ is fully known, can estimate θ_0 by Maximum Likelihood (MLE). This is the preferred method, it offers the best asymptotic precision.
- If the law $p(X, \theta_0)$ is not fully known, but we know some features of the distribution, e.g., the first two moments, we can still estimate the parameters by the quasi-MLE method.
- Alternatively, if we only know a few moments of the distribution, but not the entire pdf $p(X, \theta_0)$, we can estimate parameters by the Generalized Method of Moments (GMM).
- QMLE and GMM methods are less precise (efficient) than MLE, but they are more robust since they do not require the full knowledge of the distribution.

Outline

Math Review: Jensen's Inequality

- Jensen's inequality states that if $f(x)$ is a concave function, and $w_n \geqslant 0$, $n = 1, \ldots, N$, and $\sum_{n=1}^{N} w_n = 1$, then

$$\sum_{n=1}^{N} w_n f(x_n) \leqslant f\left(\sum_{n=1}^{N} w_n x_n\right)$$

  for any x_n, $n = 1, \ldots, N$.

- This result extends to the continuous case:

$$\int w(x) f(x)\, dx \leqslant f\left(\int w(x) x\, dx\right), \quad \text{if} \quad \int w(x)\, dx = 1, \quad w(x) \geqslant 0$$

- Example: if x is a random variable (e.g., asset return), and f a concave function (e.g., utility function), then

$$\mathrm{E}[f(x)] \leqslant f\left(\mathrm{E}[x]\right) \qquad \text{(risk aversion)}$$

Maximum Likelihood Estimator (MLE)

- IID observations x_t, $t = 1, \dots, T$ with density $p(x, \theta_0)$.
- Maximum likelihood estimation is based on the fact that for any alternative distribution density $p(x, \widetilde{\theta})$,

$$\mathrm{E}\left[\ln p(x, \widetilde{\theta})\right] \leqslant \mathrm{E}\left[\ln p(x, \theta_0)\right], \qquad \mathrm{E}[\star] = \int \star\, p(x, \theta_0)\, dx$$

- To see this, use Jensen's inequality, and equality $\int p(x, \widetilde{\theta})\, dx = 1$:

$$\mathrm{E}\left[\ln \frac{p(x_t, \widetilde{\theta})}{p(x_t, \theta_0)}\right] \leqslant \ln \mathrm{E}\left[\frac{p(x_t, \widetilde{\theta})}{p(x_t, \theta_0)}\right] = \ln \int \frac{p(x, \widetilde{\theta})}{p(x, \theta_0)} p(x, \theta_0)\, dx =$$

$$\ln \int p(x, \widetilde{\theta})\, dx = 0$$

- Estimate parameters using the sample analog of the above inequality

$$\widehat{\theta} = \arg\max_{\theta} \frac{1}{T} \sum_{t=1}^{T} \ln p(x_t, \theta) = \arg\max_{\theta} \frac{1}{T} \ln p(X, \theta))$$

Maximum Likelihood Estimator (MLE)

- Define the Likelihood function

$$L(\theta) = \ln p(X, \theta)$$

- Likelihood function treats model parameters θ variables. It treats observations X as fixed.
- We will work with the log of likelihood, $\mathcal{L}(\theta) = \ln L(\theta)$. We will often drop the "log" and simply call $\mathcal{L}$ likelihood.
- For IID observations,

$$\frac{1}{T}\mathcal{L}(\theta) = \frac{1}{T} \ln \prod_{t=1}^{T} p(x_t, \theta) = \frac{1}{T} \sum_{t=1}^{T} \ln p(x_t, \theta)$$

and therefore θ can be estimated by maximizing (log-) likelihood

MLE

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

Example: MLE for Gaussian Distribution

- IID Gaussian observations, mean μ, variance σ^2.
- The log likelihood for the sample $x_1, \ldots, x_T$ is

$$\mathcal{L}(\theta) = \ln \prod_{t=1}^{T} p(x_t, \theta) = \sum_{t=1}^{T} \ln p(x_t, \theta) = \sum_{t=1}^{T} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_t - \mu)^2}{2\sigma^2}$$

- MLE: $\widehat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$
- Optimality conditions:

$$\frac{\sum_{t=1}^{T}(x_t - \widehat{\mu})}{\widehat{\sigma}^2} = 0, \qquad -\frac{T}{\widehat{\sigma}} + \frac{\sum_{t=1}^{T}(x_t - \widehat{\mu})^2}{\widehat{\sigma}^3} = 0$$

- These are identical to the GMM conditions we have derived above!

$$\widehat{\mathrm{E}}(x_t - \widehat{\mu}) = 0, \qquad \widehat{\mathrm{E}}\left[(x_t - \widehat{\mu})^2\right] - \widehat{\sigma}^2 = 0$$

Example: Exponential Distribution

- Suppose we have T independent observations from the exponential distribution

$$p(x_t, \lambda) = \lambda \exp(-\lambda x_t)$$

- Likelihood function

$$\mathcal{L}(\lambda) = \sum_{t=1}^{T} (-\lambda x_t + \ln \lambda)$$

- First-order condition

$$\left(-\sum_{t=1}^{T} x_t \right) + \frac{T}{\widehat{\lambda}} = 0$$

implies

$$\widehat{\lambda} = \left(\frac{\sum_{t=1}^{T} x_t}{T} \right)^{-1}$$

MLE for Dependent Observations

- MLE approach works even if observations are dependent.
- Need dependence to die out quickly enough.
- Consider a time series $x_t, x_{t+1}, \ldots$ and assume that the distribution of x_{t+1} depends only on L lags: $x_t, \ldots, x_{t+1-L}$.
- Log likelihood conditional on the first L observations:

$$\mathcal{L}(\theta) = \sum_{t=L}^{T-1} \ln p(x_{t+1}|x_t, \ldots, x_{t+1-L}; \theta)$$

- θ maximizes conditional expectation of $\ln p(x_{t+1}|x_t, \ldots, x_{t-L+1}; \theta)$ and thus maximizes the (conditional) likelihood if T is large and x_t is stationary.

MLE

$$\widehat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta)$$

Outline

MLE for AR(p) Time Series

- AR(p) (**A**uto**R**egressive) time series model with IID Gaussian errors:

$$x_{t+1} = a_0 + a_1 x_t + \dots a_p x_{t+1-p} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \mathcal{N}(0, \sigma^2)$$

- Conditional on $(x_t, \dots, x_{t+1-p})$, x_{t+1} is Gaussian with mean 0 and variance σ^2.
- Construct likelihood:

$$\mathcal{L}(\theta) = \sum_{t=p}^{T-1} -\ln\sqrt{2\pi\sigma^2} - \frac{(x_{t+1} - a_0 - a_1 x_t - \dots a_p x_{t+1-p})^2}{2\sigma^2}$$

- MLE estimates of $(a_0, a_1, \dots, a_p)$ are the same as OLS:

$$\max_{\vec{a}} \mathcal{L}(\theta) \Leftrightarrow \min_{\vec{a}} \sum_{t=p}^{T-1} (x_{t+1} - a_0 - a_1 x_t - \dots a_p x_{t+1-p})^2$$

MLE for VAR(p) Time Series

- VAR(p) (**V**ector **A**uto**R**egressive) time series model with IID Gaussian errors:

$$x_{t+1} = a_0 + A_1 x_t + \dots A_p x_{t+1-p} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \mathcal{N}(0, \Sigma)$$

  where x_t and a_0 are N-dim vectors, A_n are $N \times N$ matrices, and ε_t are N-dim vectors of shocks.
- Conditional on $(x_t, \dots, x_{t+1-p})$, x_{t+1} is Gaussian with mean 0 and var-cov matrix Σ.
- Construct likelihood:

$$\mathcal{L}(\theta) = \sum_{t=p}^{T-1} -\ln \sqrt{(2\pi)^N |\Sigma|} - \frac{1}{2} \varepsilon'_{t+1} \Sigma^{-1} \varepsilon_{t+1}$$

MLE for VAR(p) Time Series

- Parameter estimation:

$$\max_{a_0,A_1,\ldots,A_p,\Sigma} \mathcal{L}(\theta) \Leftrightarrow \min_{a_0,A_1,\ldots,A_p,\Sigma} \sum_{t=p}^{T-1} \ln \sqrt{(2\pi)^N|\Sigma|} + \frac{1}{2}\varepsilon_{t+1}'\Sigma^{-1}\varepsilon_{t+1}$$

- Optimality conditions for a_0, A_1, ..., A_p:

$$\sum_t \left[x_{t-i}\varepsilon_{t+1}'\right] = 0,\ \ i = 0, 1, \ldots, p-1, \quad \sum_t \varepsilon_{t+1} = 0$$

where

$$\varepsilon_{t+1} = x_{t+1} - (a_0 + A_1 x_t + \ldots A_p x_{t+1-p})$$

- VAR coefficients can be estimated by OLS, equation by equation.
- Standard errors can also be computed for each equation separately.

Outline

MLE and Model Selection

- In practice, we often do not know the exact model.
- In some situations, MLE can be adapted to perform model selection.
- Suppose we are considering several alternative models, one of them is the correct model.
- If the sample is large enough, we can identify the correct model by comparing maximized likelihoods and penalizing them for the number of parameters they use.
- Various forms of penalties have been proposed, defining various *information criteria*.

VAR(p) Model Selection

- To build a VAR(p) model, we must decide on the order p.
- Without theoretical guidance, use an information criterion.
- Consider two most popular information criteria: Akaike (AIC) and Bayesian.
- Each criterion chooses p to maximize the log likelihood subject to a penalty for model flexibility (free parameters). Various criteria differ in the form of penalty.

AIC and BIC

- Start by specifying the maximum possible order $\overline{p}$.
- Make sure that $\overline{p}$ grows with the sample size, but not too fast:

$$\lim_{T\to\infty} \overline{p} = \infty, \quad \lim_{T\to\infty} \frac{\overline{p}}{T} = 0$$

For example, can choose $\overline{p} = \frac{1}{4}(\ln T)^2$.

- Find the optimal VAR order $p^\star$ as

$$p^\star = \arg\max_{0\leqslant p\leqslant \overline{p}} \frac{2}{T}\mathcal{L}(\theta; p) - \text{penalty}(p)$$

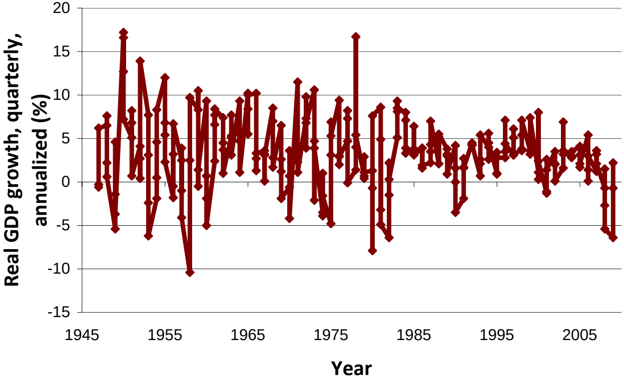
where

$$\text{penalty}(p) = \begin{cases} \text{AIC:} & \frac{2}{T} pN^2 \\ \text{BIC:} & \frac{\ln T}{T} pN^2 \end{cases}$$

- In larger samples, BIC selects lower-order models than AIC.

Example: AR(p) Model of Real GDP Growth

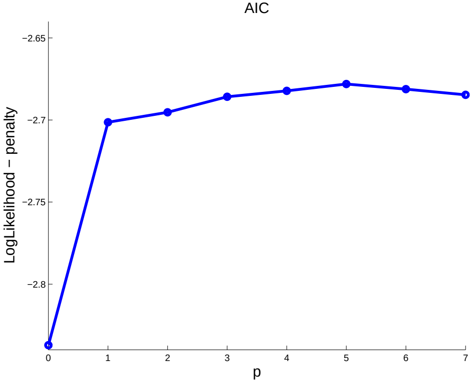
- Model quarterly seasonally adjusted GDP growth (annualized rates).
- Want to select and estimate an AR(p) model.



Source: U.S. Department of Commerce, Bureau of Economic Analysis. National Income and Product Accounts.

Example: AR(p) Model of GDP Growth

- Set $\overline{p} = 7$.

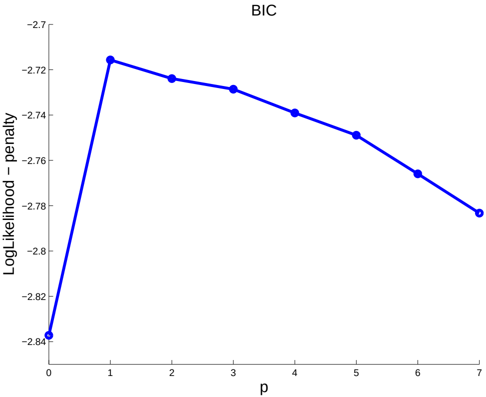


- AIC dictates $p = 5$.
- AR coefficients $a_1, \ldots, a_5$:

$$0.3185,\ 0.1409,\ -0.0759,\ -0.0600,\ -0.0904$$

Example: AR(p) Model of GDP Growth

- Set $\overline{p} = 7$.



- BIC dictates $p = 1$.
- AR coefficient $a_1 = 0.3611$.

Outline

IID Observations

- A sample of independent and identically distributed (IID) observations drawn from the distribution family with density $\phi(x;\theta_0)$:

$$X = (x_1, \ldots, x_t, \ldots, x_T)$$

- Want to estimate the N-dimensional parameter vector θ_0, .
- Consider a vector of functions $f_j(x,\theta)$ ("moments"), $\dim(f) = N$.
- Suppose we know that for any j,

$$\begin{aligned} &\mathrm{E}[f_1(x_t,\theta_0)] = \cdots = \mathrm{E}[f_N(x_t,\theta_0)] = 0, && \text{if } \theta = \theta_0 \\ &\textstyle\sum_{j=1}^N \left(\mathrm{E}[f_j(x_t,\theta)]\right)^2 > 0, && \text{if } \theta \neq \theta_0 \end{aligned} \qquad \text{(Identification)}$$

- GMM estimator $\widehat{\theta}$ of the unknown parameter θ_0 is defined by

GMM

$$\widehat{\mathrm{E}}[f(x_t,\widehat{\theta})] \equiv \frac{1}{T}\sum_{t=1}^{T} f(x_t,\widehat{\theta}) = 0$$

Example: Mean-Variance

- Suppose we have a sample from a distribution with mean μ_0 and variance σ_0^2.
- To estimate the parameter vector $\theta_0 = (\mu_0, \sigma_0)'$, $\sigma_0 \geqslant 0$, choose the functions $f_j(x, \theta)$, $j = 1, 2$:

$$f_1(x_t, \theta) = x_t - \mu$$
$$f_2(x_t, \theta) = (x_t - \mu)^2 - \sigma^2$$

- Easy to see that $\mathrm{E}[f(x, \theta_0)] = 0$.
- If $\theta \neq \theta_0$, then $\mathrm{E}[f(x, \theta)] \neq 0$ (verify).
- Parameter estimates:

$$\widehat{\mathrm{E}}(x_t) - \widehat{\mu} = 0 \Rightarrow \widehat{\mu} = \widehat{\mathrm{E}}(x_t)$$
$$\widehat{\mathrm{E}}\left[(x_t - \widehat{\mu})^2\right] - \widehat{\sigma}^2 = 0 \Rightarrow \widehat{\sigma}^2 = \widehat{\mathrm{E}}\left[(x_t - \widehat{\mu})^2\right]$$

GMM and MLE

- First-order conditions for MLE can be used as moments in GMM estimation.
- Optimality conditions for maximizing $\mathcal{L}(\theta) = \sum_{t=1}^{T} \ln p(x_t, \theta)$ are

$$\sum_{t=1}^{T} \frac{\partial \ln p(x_t, \theta)}{\partial \theta} = 0$$

- If we set $f = \partial \ln p(x, \theta)/\partial \theta$ (the *score vector*), then MLE reduces to GMM with the moment vector f.

Example: Interest Rate Model

- Interest rate model:

$$r_{t+1} = a_0 + a_1 r_t + \varepsilon_{t+1}, \quad \mathrm{E}(\varepsilon_{t+1}|r_t) = 0, \quad \mathrm{E}(\varepsilon_{t+1}^2|r_t) = b_0 + b_1 r_t$$

- Derive moment conditions for GMM.
- Note that for any function $g(r_t)$,

$$\mathrm{E}[g(r_t)\varepsilon_{t+1}] = \mathrm{E}\left[\mathrm{E}[g(r_t)\varepsilon_{t+1}|r_t]\right] = \mathrm{E}\left[g(r_t)\mathrm{E}[\varepsilon_{t+1}|r_t]\right] = 0$$

- Using $g(r_t) = 1$ and $g(r_t) = r_t$,

$$\mathrm{E}\left[(1, r_t)'(r_{t+1} - a_0 - a_1 r_t)\right] = 0$$
$$\mathrm{E}\left\{(1, r_t)'\left[(r_{t+1} - a_0 - a_1 r_t)^2 - b_0 - b_1 r_t\right]\right\} = 0$$

Example: Interest Rate Model

- GMM using the moment conditions

$$\mathrm{E}\left[(1, r_t)'(r_{t+1} - a_0 - a_1 r_t)\right] = 0$$
$$\mathrm{E}\left\{(1, r_t)'\left[(r_{t+1} - a_0 - a_1 r_t)^2 - b_0 - b_1 r_t\right]\right\} = 0$$

- (a_0, a_1) can be estimated from the first pair of moment conditions. Equivalent to OLS, ignore information about second moment.

Outline

MLE and QMLE

- Maximum likelihood estimates are optimal: they have the smallest asymptotic variance.
- When we know the distribution function $p(X, \theta)$ precisely, MLE is the most *efficient* approach.
- MLE is often a convenient way to figure out which moment conditions to impose.
- Even if the model $p(X, \theta)$ is misspecified, MLE approach may still be valid as long as the implied moment conditions are valid.
- With an incorrect model $q(X, \theta)$, MLE is a special case of GMM. GMM results apply.
- The approach of using an incorrect (typically Gaussian) likelihood function for estimation is called quasi-MLE (QMLE).

Example: QMLE for AR(p) Time Series

- AR(p) time series model with IID non-Gaussian errors:

$$x_{t+1} = a_0 + a_1 x_t + \dots a_p x_{t+1-p} + \varepsilon_{t+1}, \quad \mathrm{E}[\varepsilon_{t+1} | x_t, \dots, x_{t+1-p}] = 0$$

- Pretend errors are Gaussian to construct $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \sum_{t=p}^{T-1} - \ln \sqrt{2\pi\sigma^2} - \frac{(x_{t+1} - a_0 - a_1 x_t - \dots a_p x_{t+1-p})^2}{2\sigma^2}$$

- Optimality conditions:

$$\sum_t (x_{t-i} \varepsilon_{t+1}) = 0, \; i = 0, \dots, p-1, \quad \sum_t \varepsilon_{t+1} = 0$$

- Valid moment conditions (verify). GMM justifies QMLE.

Example: Interest Rate Model

- Interest rate model:

$$r_{t+1} = a_0 + a_1 r_t + \varepsilon_{t+1}, \quad \mathrm{E}(\varepsilon_{t+1}|r_t) = 0, \quad \mathrm{E}(\varepsilon_{t+1}^2|r_t) = b_0 + b_1 r_t$$

- GMM using the moment conditions

$$\mathrm{E}\left[(1, r_t)'(r_{t+1} - a_0 - a_1 r_t)\right] = 0$$
$$\mathrm{E}\left\{(1, r_t)'\left[(r_{t+1} - a_0 - a_1 r_t)^2 - b_0 - b_1 r_t\right]\right\} = 0$$

- (a_0, a_1) can be estimated from the first pair of moment conditions. Equivalent to OLS, ignore information about second moment.

Example: Interest Rate Model

- QMLE: treat ε_t as Gaussian $\mathcal{N}(0, b_0 + b_1 r_{t-1})$.
- Construct $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \sum_{t=1}^{T-1} - \ln \sqrt{2\pi(b_0 + b_1 r_t)} - \frac{(r_{t+1} - a_0 - a_1 r_t)^2}{2(b_0 + b_1 r_t)}$$

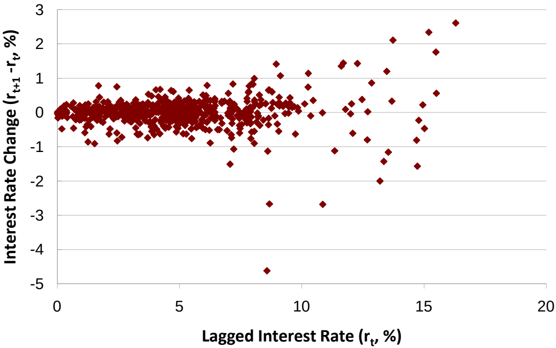
- (a_0, a_1) can no longer be estimated separately from (b_0, b_1).
- Optimality conditions for (a_0, a_1):

$$\sum_{t=1}^{T-1} (1, r_t)' \frac{(r_{t+1} - a_0 - a_1 r_t)}{b_0 + b_1 r_t} = 0$$

- This is no longer OLS, but GLS. More precise estimates of (a_0, a_1).
- Down-weight residuals with high variance.

Example: Interest Rate Model

- 3-Month Treasury Bill: secondary market rate, monthly.
- Scatter plot of interest rate changes vs lagged interest rate values.
- Higher volatility of rate changes at higher rate levels.



Source: Federal Reserve Bank of St. Louis.

Discussion

- QMLE approach helps specify moments in GMM.
- Do not use blindly, verify that the moment conditions are valid.

Key Points

- Parameter estimators, consistency.
- Likelihood function, maximum likelihood parameter estimation.
- Identification of parameters by GMM.
- QMLE. Verify the validity of QMLE by interpreting the resulting moments in GMM framework.

Readings

- Tsay, 2005, Sections 1.2.4, 2.4.2, 8.2.4.
- Cochrane, 2005, Sections 11.1, 14.1, 14.2.
- Campbell, Lo, MacKinlay, 1997, Section A.2, A.4.

15.450 Analytics of Finance
Fall 2010