Time Series Econometrics Glossary

ECON 6413–Ng

Miguel Acosta

Fall 2017

Contents

1 Time Series Basics

ARMA: Stationarity and Invertibility (For notation) $\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p$, $\theta(L) = 1 + \phi_1 L + \cdots + \phi_q L^q$. An ARMA$(p, q) = \alpha(L)y_t = \theta(L)u_t$.

- AR(p) is stationary if the roots of the characteristic polynomial $z_p - \alpha_1 z^{p-1} - \cdots - \alpha_p = 0$ are inside the unit circle; this can also be written as the roots of $1 - \alpha_1 L - \cdots - \alpha_p L^p = 0$ are outside the unit circle.[1]
- MA(q) are always stationary.
- AR(p) are always invertible.
- MA(q) is invertible if the roots of $\theta(z) = 0$ are inside the unit circle, or the roots of $\theta(L) = 0$ are outside.

Ergodicity

- **Stationary Ergodic** A process $\{y_t\}$ is stationary ergodic the process is "asymptotically independent" of itself. That is, for any functions—$f(y_t, \ldots, y_{t+k})$ and $g(y_{t+T}, \ldots, y_{t+T+\ell})$—which are functions of two subsamples, they become independent as the subsamples grow apart; that is, $\mathbb{E}[f \cdot g] = \mathbb{E}[f]\,\mathbb{E}[g]$ as $T \to \infty$. As an example, y_t is stationary ergodic if $\gamma_k \to 0$ as $k \to \infty$.
- **Ergodic for the Mean** The time average converges to the ensemble average as $T \to \infty$

Impulse Response Function $\frac{\partial\, E[y_{t+k}|e_t \mathcal{F}_{t-1}]}{\partial\, e_t}$ where e_t are the innovations in an MA representation.

Long- and Short-Run Variance The short-run variance for a series is γ_0. The long-run variance is

$$\lim_{T \to \infty} T \operatorname{var}(\overline{y}) = \gamma_0 + 2 \sum_{k=1}^{\infty} \gamma_k \equiv \omega^2$$

which is the quantity used for the LLN for time series; $\sqrt{T}\,(\overline{y} - \mu) \sim N(0, \omega^2)$.

[1] Divide by L^p...

Observational Equivalence Two processes with the same mean, variance, and autocovariances are second order observationally equivalent. Arises with e.g. MA(1) processes in which there are two possible DGPs that are observationally equivalent, so we restrict ourselves to cases with MA parameter (absolutely) less than 1.

Martingale Difference Sequence The mean of u_t cannot be forecasted; higher moments may be. Technically, $\mathbb{E}(u_t \mid \mathcal{F}_{t-1}) = 0$.

Stationarity A random variable y_t is covariance stationary if the first two unconditional moments (mean, variance, autocovariances) do not depend on t. Strict stationarity means that the distribution is constant over time—this implies that all moments and functions of y_t are independent of t.

Wold Decomposition Any covariance stationary process can be represented as the sum of two mutually uncorrelated processes, where one is a linear forecast (deterministic) and the other is stochastic mean-zero forecast error. This implies that the process has a unique MA(∞) representation.

White Noise $\mathbb{E}(u_t) = 0$ and $\mathbb{E}(u_t u_{t+j}) = 0 \forall j \neq 0$; this means that u is linearly unpredictable—lags of u cannot help predict u. (Independent White Noise means that $u_t \sim$ iid $(0, \sigma^2)$.

2 Estimation & Model Selection

Setup: a log-likelihood, its gradient, and its Hessian, respectively:

$$\ln L_t(\theta) \equiv \frac{1}{T}\left\{\ln f(y_1 \mid x_x, \theta) + \sum_{t=2}^{T} \ln f(y_t \mid \{y_{t-p}\}_{p=1}^{t-1}, \{x_{t-p}\}_{p=0}^{t-1})\right\} \quad G_T(\theta) \equiv \frac{\partial \ln L_t(\theta)}{\partial \theta} \quad H_T(\theta) \equiv \frac{\partial^2 \ln L_t(\theta)}{\partial \theta \partial \theta'}$$

Identification (and, thus, consistent estimability) depends crucially on the information matrix ($\lim_{T\to\infty} \mathbb{E}\, H_t$) is invertible.

Optimization By first order approximation, $\widehat{\theta} = \theta_0 + H_T^{-1} G_T(\theta_0)$. Trying to maximize the likelihood, which is equivalent to trying to zero out this equation. There are several methods. **Simulated annealing** is interesting because it has some probability of not getting stuck in a valley of the likelihood.

GMM/Minimum Distance We covered GMM last year; the key in time series data is to pick *good* moments that take into account the time series properties—for example, autocovariances.

Simulation Estimation If you can solve it, you can simulate it. And if you can simulate it, you can estimate it. Pick some moment, ψ. Draw several $S \times T$ errors, and hold these constant. Add the errors to your model to get S "samples", and take the average. This is your $\overline{g}(\psi)$. Vary ψ to maximize $\overline{G}' W \overline{G}$.

Indirect Inference Write a model that is easier to estimate. Figure out the mapping between this **auxiliary model** and your model. Estimate the auxiliary model. Use the mapping to back out the estimates for your original model

Granger Causality z_k does not Granger cause y_t if the lags of z_k do not have predictive power for y_t. Just include them in the model and run standard tests (LM, LR, etc.).

Structural Breaks Run regression, for each t, with a dummy at t interacted with whatever you want. Run Wald-test. Take largest. Largest is indiciative of structural break.

Diagnostics Serially uncorrelated errors, non-linearities, ARCH. We didn't discuss these much.

Model selection Considering two models for a time series, M_1 and M_2, we want want a consistent procedure for distinguishing: $P(\widehat{M} = M_i \mid M_i) \to 1$. Some approaches are

- Test statistic: pick some statistic, say Wald, then choose M_1 if it gives you a satisfactory value. Otherwise, move to other. Could have Type I error.
- R^2 Compare R^2's. Do not use this.
- Information criteria: tradeoff between goodness of fit (sum of squared errors) and complexity of the model (number of parameters). Different penalties possible. Make sure you always use same sample if there is trimming. Serena has a survey paper on this.
- Encompassing and nested models...

Forecasting Univariate; Kolmogorov-Weiner; Companion matrix;

3 VAR

Representation A VAR(p) is just an AR(p) where, instead of a scalar, we have a vector. Any finite order VAR can be put in companion form—that is, represented as a VAR(1)—using some creative matrix operators.

Stability A VAR is stable if the companion has all eigenvalues less than 1 in modulus. Alternatively, if $\det(I_n - A_1 z - \cdots - A_p z^p) \neq 0$ for any $z \in \mathbb{C}$ less than one in modulus.

Estimation Generally, GLS would be efficient. OLS is equivalent to GLS if the regressor matrix is the same for all equations—so, we always do this (that is, use all lags of all variables in all equations). Estimation can then be done by OLS, one at a time, or using some fancy vector OLS.

Inference The estimated coefficients have analytically tractable asymptotic variance—but we tend not to use them. Instead, bootstrapping is the usual way. **Runkle** (drawing errors and iterating them through the system) is in the notes. **Wild bootstrap** is also common; there is also a parametric bootstrap. Kilian's **bootstrap after the bootstrap** has you bootstrap after you bias-correct your estimates.

Granger Causality Is easy to read off from the coefficients using a Wald test.

Model Selection A few information criteria tests which reward goodness of fit, but penalize model complexity. Other tests include LR, sequential testing, and Portmonteau tests.

MA Representation Useful for **forecasting**—using previous values to forecast future—*IRF*—looking at how shocks to one variable affect other—and **variance decomposition**—how much of the variance in each variable can be explained by shocks to the others.

FAVAR You can include the principle components of some other variable (called "factors") as long as you use a lot of variables when finding the principle components.

Identification We have more variables to identify than we have estimated coefficients, so we need to make some restrictions. The structural and reduced form models, respectively, are:

$$\begin{aligned} B_0 y_t &= B_1 y_{t-1} + \cdots + B_p y_{t-p} + R v_t \\ y_t &= A_1 y_{t-1} + \cdots + A_p y_{t-p} + e_t \end{aligned}$$

Typically, we put restrictions on the matrices that determine simultaneous impacts—that is, we restrict which *shocks* affect which variables contemporaneously (restricting R), or which variables affect which variables contemporaneously (restricting B_0). Methods (detailed in the notes) include:

- **Recursive** identification (ordering) puts restrictions on B_0 and R; assumes that variable n is not contemporaneously affected by variable $n+1$.
- **Non-recursive ordering** put zeros in B_0 and R as you like, and estimate with minimum distance. *These exclusion restrictions can be cast as IV.*
- **Long-run** restrictions place restrictions on the long-run impact of shocks. That is, they are driven by theoretical equilibrium conditions. Drawbacks: (1) requires LR variance to be precisely estimated, which isn't easy with persistent data (2) identified up to sign.
- **Narrative approach**
- **Heteroskedasticity** put restrictions on the covariance matrix of the reduced-form shocks by assuming that there are two regimes in which the variance differs, but the parameters stay unchanged.
- **Sign restrictions** place restrictions on the signs of the impulse responses. Only results on **set-identification**.
- **External IV** Isolate to focus on one shock only; use external shock as instruments (see replication exercise in problem set 3).

Fundamentalness In the case of, for example, news shocks, then structural shocks can't be recovered as a function of the past alone. In fact, they depend on the entire future too!

4 Kalman Filter

The point of the Kalman filter is the following—many models can be written in state-space form. The likelihood function for this data can be easily written as a function of the products of the Kalman filtering process. Thus, estimation of the original model by MLE is relatively easy once the model can be put in state-space form.

Recursive OLS Estimate β_t in an OLS regression by using all data up to time t; do this for all t. Nice feature is that the errors are **mutually uncorrelated**. Also, Looking at graph of β_t is a way to gauge **parameter instability** (CUSUM, CUMSUMQ).

Motivation for Kalman Filter The likelihood can be expressed in terms of prediction error and prediction error variance, both of which are byproducts of the Kalman Filter.

State-Space Representation There is a measurement equation, which expresses variables you can see as a function of the state variable. There is also a transition equation, which tells you how the state evolves from period to period. There are shocks to both systems (measurement error in the first, and shocks to the state in the latter).

Examples of state-space models Dynamic factor (some underlying growth rate, for example); time varying parameters; ARMA. Of course there are many others.

Kalman Filter Once you have a system in state-space form, start by initializing a value for the state (a few ways to do this). Perform prediction and updating of the state, one period at a time (similar to recursive OLS). At each point in time the optimal estimator of the state is a weighted average of the forecast for the current state using last period's state and the Kalman gain, which includes information about how bad your estimate was. You can go back and "smooth" using all of the information for the whole sample. Smoothing vs. filtering depends on the exercise you're performing.

5 Bayesian and Quasi-Bayesian Estimation

Key: estimates are combinations of prior distribution and data. As the sample size increase, the contribution of the prior decreases.

Priors Frequentists use priors too—some examples are the ridge estimator and restricted OLS. For Bayesians, the prior is some belief about the parameter of interest, before seeing the data.

Bayes Rule For θ a parameter of interest and y the data,

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} \qquad \equiv \qquad \text{posterior} = \frac{\text{likelihood } \times \text{ prior}}{p(\text{data})} \tag{1}$$

Kernels The purpose of the denominator in (1) is to ensure that the posterior integrates to 1—it isn't of interest and doesn't depend on θ. So, we typically drop it and consider only the **Kernel**:

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta).$$

Sufficiency and Likelihood Principles Inference about θ depends on the sample y only through sufficient statistics $T(y)$. The likelihood contains all the information about the data. But, when the likelihood is flat (or has two maxima), then using a tiny bit of prior information can guide us.

Conjugate Priors are useful; this is the case when the posterior distribution has the same parametric form as the prior. I think this is important because when we have to estimate things, we have to iterate and it would be hard to iterate if the functional form was changing.

Computing Posteriors Find the sufficient statistic and write the kernel in terms of it. Then, find the conjugate prior, then solve for out the posterior by just multiplying the prior by the likelihood. Along the way, drop anything that doesn't depend on θ. (Many examples worked out in the notes.)

Diffuse Prior Assuming mean and variance are independent, this is

$$p(\mu, \sigma^2) = c\frac{1}{\sigma^2} \propto \frac{1}{\sigma^2}$$

The latter term is a typical diffuse prior for variance; equivalent to putting a uniform prior on $\log \sigma^2$, then doing the Jacobian transformation.

Linear model Same arguments, since $\widehat{\beta}$ is just a sample mean. *Inverse* Γ *is* χ^2 *with different parameterization.* g-prior useful.

Semi-conjugate prior If, for examplle, we put a conjugate prior on β in a linear model, but diffuse on σ^2, then it's hard to find analytic solution. Need Gibbs sampling. Remember,

BVAR Prior shrinks the effects as lags increase; makes sense if you have stationary data!

EM For $\theta = (\phi, \gamma)$, EM finds the mode of the marginal posterior of $p(\phi \mid y)$ by averaging over γ and iterating. There's a very instructive example with missing data in the notes—fill in missing values by most recent $\widehat{\beta}$, then recompute $\widehat{\beta}$, and iterate. g

Quasi-Bayesian Estimation Taking a GMM objective function and normalizing it can get you a density which you can then to Bayesian estimation on. And, Bayesian inference gives you valid frequentist inference (which should not surprise us, since the data always dominates the prior in the Bayesian world).

6 Unit Roots

Superconsistency Unit roots have super-consistency (converge at a faster rate). Tradeoff—non-standard asymptotics. Much of this lecture is concerned with how transform the estimated equation in order to identify the unit-root, while being able to use some asymptotics that have been figured out for us. It is ok to estimate equations where different parameters converge at different rates.

Terminology

- A Trend is a persistent long-term movement of a variable over time (some are deterministic, e.g., time trend).
- The following are all the same: stochastic trend, autoregressive unit root, I(1), random walk.
- Difference stationary means that the first difference is stationary

Beware If you difference a stationary variable, you induce a moving-average unit root!

Dickey-Fuller Tests See the notes—very clear.

Cointegration

- Variables are cointegrated if they are both I(1) but some quasi-difference of them is stationary.
- These include error correction mechanisms that bring us back to a long-run trend.
- We can think of conitegrated variables in a factor structure where the factor is I(1). In the growth model, A is the factor. Leads to rank-reduction.
- Every cointegrated model has 3 representations (1) an AR (sum of coefficients are reduced-rank), an MA (same), and a VECM.

THE "PRACTICAL RULES" PORTION OF THE NOTES ARE VERY USEFUL

7 The Spectrum

Spectral Decomposition Our time series can be represented as sums of J sines and cosines with different periodicities. Taking $J \to \infty$, the spectral representation of a time

Frequency The letter ω in the notes denotes the frequency. As $\omega \to 0$, we are discussing low-frequency movements (growth people work on this). As $\omega \to \pi$, we get into high-frequency (finance) time series. In the middle is macro (between 6 and 32 quarters).

Long-Run Variance Is the spectrum of x at frequency 0.

Interesting Things Taking a first-difference removes the low-frequency components of a time-series. Kuznets found that you can generate interesting-looking time series using i.i.d. data (which should have a flat spectrum).

Filtering Filtering data changes the spectrum, and can thus lead you to spurious-looking results.

Unit roots Take infinite amount of time to get back to the same point—infinite mass at 0-frequency. Take difference to remove all 0-frequency movements.