Peter C.B. Phillips' contribution to the theory (and practice) of model choice

Model choice: Given are models with (nested) parameter spaces which are nested.

Examples:

- AR, MA processes of different order
- Linear models with different order
- Sieve-type approximations

Maximum-Likelihood estimation: Chooses the "highest" order

Traditional recipe: Subtract "penalty term" from logarithm of likelihood before maximization

T = sample size, k = dimension of parameter space

AIC (Akaike): k

Hannan-Quinn: $k\log(\log(T))$

BIC (Shwartz): $k\log(T)$

Rissanen: Different justification, similar to BIC

All criteria were developed with stationary processes in sight.

Simple "nonstationary" problem:

Let u_t be i.i.d. $G(0,1)$

(The assumption of the variance to be one allows us to work without a lot of σ^2)

Assume that

$$y_t = x_t'\beta + u_t,$$

and that we want to predict y_t but do not know β:

We would have to *estimate* β by - say OLS.

Prediction error for y_t:

$$u_{t+1} + x'_{t+1}\widehat{\beta}_t.$$

Assume x_t to be independent of β: $X_t = stack\ of\ x'_1, ..$

Conditional expectation of squared estimation error:

$$(1 + x'_{t+1}(X'_t X_t)^{-1} x_{t+1})$$

derivative of logarithm of determinant:

$$d(\log\det\Sigma) = tr(\Sigma^{-1}d\Sigma)$$

Hence

$$\begin{aligned}
&\log\det(X'_{t+1}X_{t+1}) - \log\det(X'_tX_t) \\
\approx\ & tr\Big((X'_tX_t)^{-1}(x_{t+1}x'_{t+1})\Big) \\
=\ & x'_{t+1}(X'_tX_t)^{-1}x_{t+1}
\end{aligned}$$

Therefore accumulated error due to estimation:

$$\sum x'_{t+1}(X'_tX_t)^{-1}x_{t+1} \approx \log\det(X'_TX_T) - const$$

Stylized fact: Let $k = \dim x_t$

$$\log \det(X_T' X_T)$$

is $\approx k \log T$ if $(X_T' X_T)/T \to const$

Nonstationary x_t generate different rates

Therefore: Adding "unnecessary" regressors increases prediction error at different rates!

"Life is different for nonstationary regressors"

Stylized fact 2:

Estimation of $k-1$ dimensional model: $\beta^{(k)}, x_t^{(k)}$: k-th component of β, x_t

$\beta^{(k-1)}, x_t^{(k-1)}, X_t^{(k-1)}$ first $k-1$ components

$$y_t = \left(y_t - \beta^{(k)} x_t^{(k)}\right) + \beta^{(k)} x_t^{(k)}$$

First term on rhs: Linear combination of u_t and first $k-1$ components

Hence: Conditional expectation of squared prediction error

$$1 + x_{t+1}^{(k-1)\prime} \left(X_t^{(k-1)\prime} X_t^{(k-1)}\right)^{-1} x_{t+1}^{(k-1)} + \beta^{(k)^2} r_t^2$$

where r_t is the residual from a regression of $x_t^{(k)}$ on the first $(k-1)$ regressors.

Therefore sum of prediction errors due to estimation and mis-specification:

$$\log\det(X_T^{(k-1)\prime} X_T^{(k-1)}) + \beta^{(k)2} \sum r_t^2$$

Difference to error if one estimates a model with k regressors:

$$-\log\det(X_T^{(k)\prime} X_T^{(k)}) + \log\det(X_T^{(k-1)\prime} X_T^{(k-1)}) + \beta^{(k)2} \sum r_t^2$$

Geometry of OLS and a little algebra:

$$-\log\det(X_T^{(k)\prime}X_T^{(k)}) + \log\det(X_T^{(k-1)\prime}X_T^{(k-1)}) \approx -\log(\sum r_t^2)$$

Assume $\beta^{(k)}$ to be "small". If

$$\beta^{(k)^2}\sum r_t^2 - \log(\sum r_t^2) < 0$$

it is better to use the mis-specified, lower dimensional model.

Testing for $\beta^{(k)} = 0$ (F-test):

Noncentrality parameter $\beta^{(k)^2} \sum r_t^2$: If it is $>> 1$, but remains $O(1)$:

Test will reject $\beta^{(k)} = 0$ with high probability

On the other hand (if $\beta^{(k)^2} \sum r_t^2 = O(1)$):

$$\beta^{(k)^2} \sum r_t^2 - \log(\sum r_t^2) \to -\infty$$

Testing may reveal "true" model - but this may not be the optimal one!

PIC: minimize

$$\widehat{\sigma}_k^2 + \log\det(X'X)$$

will - as long as "noncentrality parameter" remains $O(1)$ - choose lowerdimensional model!

Comparison $(k-1)$ to k regressors: Lose $O(1)$ in the first term,

but win

$$\log(\sum r_t^2) \approx C.\log T$$

in the second term!

Peter’s main programmatic paper

Econometric Model Determination
Econometrica, Vol. 64, No. 4. (Jul., 1996), pp. 763-812.

(Fisher-Shultz lecture at the ESEM 1994)

Main results:

- Nonstationary models are different in two ways:
  - Inclusion of different variables have radically different consequences
  - Various Concepts (Information matrix ?) are not straightforward any more
- Both classical and Bayesian arguments can justify the same criterion
- Information theory is a useful tool, even in the nonstationary situation
- The result of statistical inference is a model, which is necessarily different from the "real" data generating process

Many (asymptotic) estimation problems are similar to

standard regression:

$X'X$ gets replaced by (empirical) information matrix,

x_t are scores of log-likelihood,.

Prediction by using models: Conditional Probability measures - use the "past" to construct conditional probability for future

Measure of distance: relative entropy (instead variance)

OLS estimator in our exercise: One needs to show that loss $(\log(\det(..)))$ holds for arbitrary estimators (true for parameters except a set of Lesbesue-measure zero).

Additional Result: Bayesian and plug-in of many ML-estimators give the same "model" = conditional probability measure!

Transformation to Bayesian model (exponential martingale): Relation to Khmaladze-transformation?

One of the few examples where econometric techniques were used in economic theory: A. Sandroni, Blume/Easley

Example: Choice of number of cointegrating relations:

Model selection in partially nonstationary vector autoregressive processes with reduced rank structure

John C. Chao, Peter C.B. Phillips

Journal of Econometrics 91, (1999) 227-271

Computes PIC-criterion for e.g. the number of cointegrating relations.

Evaluation:

Testing for cointegrating rank via model selection: evidence from 165 data sets

Badi H. Baltagi · ZijunWang

Empirical Economics (2007)

Conclusion of paper: "*Together with the simulation results in the literature, the overall empirical evidence presented here indicates that the model selection approach (especially SIC and PIC) can be a useful complement to the widely used parametric tests in cointegration analysis for applied researchers.*"

Another advantage of PIC

Old problem (J. Stock, sometimes in the last century..):

Cointegration relies on the fact that all of the processes involved are nonstationary.

Therefore one tests for unit roots.

All of the unit-root-tests, are, however, relatively weak.

So it might be the case that the "true" data-generating process is only near-unit-root, and therefore stationary.

Can we define cointegration, can we do inference?

VAR Model:

$$\Delta y_t = F y_{t-1} + \ldots + u_t$$

Cointegration

F is singular

Alternative y_t is "near unit root" and "cointegrated"

$$\Delta y_t = (F + \frac{1}{T} A) y_{t-1} + \ldots + u_t$$

and

F is singular

Tests for cointegration will get cranky in this kind of situation

PIC: will not be bothered by additional term:

When "noncentrality parameter" for test remains $O(1)$,

PIC decides for the simpler model:

Defines "Cointegration" in this situation.

When sample size increases, PIC will decide for the complex model,

against singular F

Stronger form of Occam’s razor: A new parameter should only be introduced, if there is enough information available so that estimation error has no detrimental effects on the model.