NBER Summer Institute

What's New in Econometrics – Time Series

Lecture 11

July 16, 2008

Forecasting and Macro Modeling

with Many Predictors, Part I

Outline

Lecture 11

1) Why Might You Want To Use Hundreds of Series?
2) Dimensionality: From Curse to Blessing
3) Dynamic Factor Models: Specification and Estimation

Lecture 12

4) Other High-Dimensional Forecasting Methods
5) Empirical Performance of High-Dimensional Methods
6) SVARs with Factors: FAVAR
7) Factors as Instruments
8) DSGEs and Factor Models

1) Why Might You Want To Use Hundreds of Series?

A theme of these lectures has been the challenge of working with limited information (problems of identification) in macro time series. But all the work until now has focused on models with relatively few variables. In fact, however, thousands of economic time series are available on line in real time. Can these be used for economic monitoring and forecasting? For estimation of single and multiple equation models?

This is a radical proposal!

- not your "principle of parsimony"!
- VARs with 6 variables and 4 lags have $4 \times 6^2 = 144$ coefficients (plus variances)

Why use hundreds of series, ctd.

We will consider four specific problems in which more information would be most welcome:

1. Economic monitoring ("nowcasting") and forecasting
   - can we move from small models with forecasts adjusted by judgmental use of additional information, to a more scientific system that incorporates as much quantitative information as possible?
2. SVARs using more information
   - so innovations span the space of shocks
3. IV estimation
   - more information might produce stronger instruments
4. DSGE estimation
   - more information might produce stronger identification

Why use hundreds of series, ctd.

It turns out that dynamic factor models (Geweke (1977), Sargent and Sims (1977)) have proven very useful in this research program

- The greatest amount of experience to date with DFMs is for forecasting. DFMs are in use for real-time monitoring and forecasting (e.g. CFNAI (Federal Reserve Bank of Chicago), Giannone, Reichlin, and Small (2008), Aruoba, Diebold, and Scotti (2008)
- Other promising applications
  - SVARs: Bernanke, Boivin, and Eliasz's (2005) FAVAR
  - DSGEs: Boivin and Giannoni (2006b)

In a broader sense, the move of empirical macro to use much larger data sets is consistent with developments in other scientific areas – mainly experimental sciences (especially life sciences/genomics) but also some observational sciences (astrophysics).

Outline of the next two lectures

1) Why Might You Want To Use Hundreds of Series?
2) Dimensionality: From Curse to Blessing
3) Dynamic Factor Models: Specification and Estimation
4) Other High-Dimensional Forecasting Methods
5) Empirical Performance of High-Dimensional Methods
6) SVARs with Factors: FAVAR
7) Factors as Instruments
8) DSGEs and Factor Models

2) Dimensionality: From Curse to Blessing

The curse part:

- A VAR with 200 variables and 6 lags has 240,000 coefficients, and another 20,100 variance parameters.
- This has really bad consequences for OLS. Here is a short calculation: Consider the regression model,

$$Y_{t+1} = \delta' P_t + \varepsilon_{t+1,}\ t = 1,\ldots,\ T,$$

$P_t = n$ orthonormal predictors so $\mathbf{P}'\mathbf{P}/T = I_n$

P_t strictly exogenous, ε_{t+1} i.i.d. $N(0,\sigma_\varepsilon^2)$

Consider quadratic forecast loss function, $L(Y_{T+1},\tilde{Y}_{T+1|t}) = (Y_{T+1} - \tilde{Y}_{T+1|t})^2$

What is the forecast risk (expected loss) of OLS?

OLS with too many regressors

$$Y_{t+1} = \delta' P_t + \varepsilon_{t+1,}\ t = 1,\ldots,\ T$$

Recall from Lecture 8 that the (frequentist) risk is the expected loss:

forecast loss: $L(Y_{T+1}, \tilde{Y}_{T+1|t}) = (Y_{T+1} - \tilde{Y}_{T+1|t})^2$

forecast risk = $EL(Y_{T+1}, \tilde{Y}_{T+1|t}) = E(Y_{T+1} - \tilde{Y}_{T+1|t})^2$

$$= E[(\tilde{\delta} - \delta)P_T + \varepsilon_{T+1}]^2$$
$$= E[(\tilde{\delta} - \delta)' P_T\, P_T'\, (\tilde{\delta} - \delta)] + \sigma^2$$
$$\approx E[(\tilde{\delta} - \delta)'(\tilde{\delta} - \delta)] + \sigma_\varepsilon^2 \text{ because } P_t \text{ is orthonormal}$$
$$= R(\tilde{\delta}, \delta) + \sigma_\varepsilon^2$$

where

$$R(\tilde{\delta}, \delta) = E[(\tilde{\delta} - \delta)'(\tilde{\delta} - \delta)]$$
$$= E\text{tr}[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)']$$

is the frequentist estimation risk, often call the "trace MSE" risk because $\text{tr}[(\tilde{\delta} - \delta)(\tilde{\delta} - \delta)']$ is the trace MSE loss (trace of the MSE matrix of $\tilde{\delta}$)

OLS with too many regressors, ctd

$$EL(Y_{T+1},\tilde{Y}_{T+1|t}) = R(\tilde{\delta},\delta) + \sigma_\varepsilon^2,\ R(\tilde{\delta},\delta) = E\text{tr}[(\tilde{\delta}-\delta)(\tilde{\delta}-\delta)']$$

- If you knew δ, then you would use $\tilde{\delta} = \delta$, in which case $R(\tilde{\delta},\delta) = 0$ and $EL(Y_{T+1},\tilde{Y}_{T+1|t}) = \sigma_\varepsilon^2$
- If $\tilde{\delta} \xrightarrow{p} \delta$ then $R(\tilde{\delta},\delta) \to 0$, so the forecast risk would $\to 0$ and the forecast would be first-order efficient (there would be second order risk because of estimation error).
- But if n is large, OLS is not first-order efficient: because **P** is strictly exogenous and ε_t is i.i.d. N(0,σ_e^2),

$$\tilde{\delta} - \delta \sim N\left(0, \left(\frac{\mathbf{P}'\mathbf{P}}{T}\right)^{-1}\sigma_\varepsilon^2\right) = N\left(0, I_n\frac{\sigma_\varepsilon^2}{T}\right)$$

so $R(\tilde{\delta},\delta) = E\text{tr}[(\tilde{\delta}-\delta)(\tilde{\delta}-\delta)'] = E\text{tr}[I_n\sigma_e^2/T] = \frac{n}{T}\sigma_e^2$. Thus:

forecast risk of OLS = $R(\tilde{\delta},\delta) + \sigma_\varepsilon^2 = (1+\kappa)\sigma_e^2$, where $\kappa = n/T$

OLS with too many regressors, ctd

OLS forecast risk: $EL(Y_{T+1}, \tilde{Y}_{T+1|t}) = (1+\kappa)\sigma_e^2 > \sigma_e^2$, where $\kappa = n/T$

- If $\kappa \approx 0$, OLS is nearly first order efficient. Parsimony!
- If n/T is large, this result provides a theory to support your intuition: OLS doesn't achieve first-order forecast efficiency
- Moreover, if $n \geq 3$, OLS isn't admissible under trace MSE loss (Stein (1955)): there exists an estimator $\tilde{\tilde{\delta}}$ with frequentist risk $R(\tilde{\tilde{\delta}}, \delta)$ that dominate OLS (risk at least as good as OLS for some δ, and no worse for all δ). James and Stein (1960) constructed a shrinkage estimator that dominates OLS (does better than OLS near $\delta = 0$).
- Things that don't achieve first order forecast efficiency:
  - throwing out all but a few regressors (throw away information!)
  - keeping only the statistically significant regressors
  - choosing regressors by information criteria (AIC or BIC)

The blessing of dimensionality, part 1

We can do better than OLS. Recall the forecasting risk is $R(\tilde{\delta},\delta) + \sigma_\varepsilon^2$. We can do nothing about σ_ε^2, but $R(\tilde{\delta},\delta)$ depends on the estimator used so the choice of estimator can reduce $R(\tilde{\delta},\delta)$.

Setup

- Adopt a local nesting in which $\delta_i = d_i/\sqrt{T}$ (else there would be an $R^2 = 1$ forecasting regression in the limit – this keeps the regression *ESS* from exploding as we let $T \to \infty$)
- Let $\{d_i\}$ have the empirical cdf G_n (suppose you observed $\{d_i\}$ – just construct the empirical cdf of $\{d_i\}$, that is G_n)
- Consider only estimators that (sensibly) produce the same forecast no matter how you order the regressors ("permutation equivariant")

The blessing of dimensionality, part 1, ctd.

Frequentist risk for permutation equivariant estimators:

$$R(\tilde{\delta},\delta) = \sum_{i=1}^{n} E(\tilde{\delta}_i - \delta_i)^2 \qquad \text{(trace MSE loss)}$$

$$= \left(\frac{n}{T}\right) n^{-1} \sum_{i=1}^{n} E(\tilde{d}_i - d_i)^2 \qquad \text{(because } \delta_i = d_i/\sqrt{T}\text{)}$$

$$= \kappa \int E(\tilde{d} - d)^2 \, dG_n(d) \qquad \text{(permutation equivariance \& cdf } G_n\text{)}$$

$$= \kappa r_{G_n}(\tilde{d}) \qquad \text{(Bayes risk* of estimator } \tilde{d} \text{ wrt } G_n\text{)}$$

where $\kappa = n/T$. *Thus the frequentist risk for permutation equivariant estimators is the Bayes risk with respect to the empirical cdf of the d's, G_n.*

*Recall from Lecture 8 that the Bayes risk $R_{G_n}(\tilde{d})$ is the expectation of the frequentist risk, with respect to a prior distribution

The blessing of dimensionality, part 1, ctd.

$$R(\tilde{\delta},\delta) = \sum_{i=1}^{n} E(\tilde{\delta}_i - \delta_i)^2 = c\int E(\tilde{d} - d)^2 \, dG_n(d) = r_{G_n}(\tilde{d})$$

This expresses a deep link between Bayes and frequentist inference.

- If you knew G_n, then you could compute the Bayes estimator w.r.t. G_n. This minimizes the Bayes risk over all estimators.
- But $R_{G_n}(\tilde{d}) = R(\tilde{\delta},\delta)$, so if $R_{G_n}(\tilde{d})$ is minimized, so is $R(\tilde{\delta},\delta)$.
- Thus, the Bayes estimator using the "prior" G_n is the optimal frequentist estimator
- How can one prior be better than another? From a subjectivist Bayes perspective, it can't be – but think about forecasting using a dogmatic prior that (say) VAR coefficients are always zero. You might have that opinion, but it would be a dumb opinion (would forecast poorly).

The blessing of dimensionality, part 1, ctd.

The ***Empirical Bayes estimator*** uses the data to pick the prior

Books on Empirical Bayes: Maritz and Lwin (1989), Carlin and Louis (1996), and Lehmann and Casella (1998, Section 4.6).

Frequentist: $\min_{\tilde{\delta}} r_{G_n}(\tilde{d}) = \kappa \int E(\tilde{d} - d)^2 \, dG_n(d)$ cdf of d_i

Bayes: $\min_{\tilde{\delta}} r_{G}(\tilde{d}) = \kappa \int E(\tilde{d} - d)^2 \, dG(d)$ subjective prior

Empirical Bayes: $\min_{\tilde{\delta}} r_{\hat{G}}(\tilde{d}) = \kappa \int E(\tilde{d} - d)^2 \, d\hat{G}(d)$ estimated "prior"

- Under technical conditions, the Empirical Bayes estimator is asymptotically admissible and asy. optimal (Robbins (1964))
- James-Stein (1960) is Empirical Bayes (Efron and Morris (1973))
- EB has certain minimax properties (Zhang (2003, 2005))
- $\hat{G}$ can be nonparametric or parametric
- asymptotically, EB is minimum risk equivariant (Edelman (1988), Knox, Stock, Watson (2001) for regression)

The blessing of dimensionality, part 1, ctd.

These are great results, but they have not been proven in time series contexts with predetermined predictors. However they are still useful guides. They tell us:

- Shrinkage (Bayes) methods can produce good forecasts (from a frequentist risk perspective) with many predictors
- Bayes methods with tuned (estimated) parameters are particularly appealing
- Forecasts using many predictors can outperform forecasts using no or only a few predictors.
- AIC, etc is not the optimal thing to do.

We will return to these methods – some theory, some empirical results.

The blessing of dimensionality, part 2

The second example is estimation of factors in a dynamic factor model (Geweke (1977), Sargent and Sims (1977)). Suppose the n variables in X_t are related to some unobserved factors F_t, which evolve according to a time series process:

$$X_t = \Lambda F_t + e_t$$

$$F_t = \Phi(\mathrm{L})F_{t-1} + G\eta_{t,}$$

If the factors were observed they could be very useful for forecasting, but they aren't observed.

The original approach to this problem (Engle and Watson (1981), Stock and Watson (1989, 1991), Sargent (1989), Quah and Sargent (1993)) was to fit the two equations above by ML using the Kalman filter. But the proliferation of parameters and computational limitations of ML in high dimensions limited this approach to small n.

The blessing of dimensionality, part 2, ctd.

How could many series be a blessing? Geweke (discussing Quah and Sargent (1993)) suggested that many series could improve estimates of F_t considerably.

An example following Forni and Reichlin (1998). Suppose F_t is scalar so Λ is a vector with elements λ_i so

$$X_{it} = \lambda_i f_t + e_{it}$$

Then $$\frac{1}{n}\sum_{i=1}^{n} X_{it} = \frac{1}{n}\sum_{i=1}^{n}\left(\lambda_i F_t + e_{it}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)F_t + \frac{1}{n}\sum_{i=1}^{n} e_{it}$$

If the errors u_{it} have limited dependence across series, then as n gets large,

$$\frac{1}{n}\sum_{i=1}^{n} X_{it} \xrightarrow{p} \bar{\lambda} F_t$$

In this special case, a very easy nonparametric estimate (the cross-sectional average) is able to recover F_t – as long as n is large!

From curse to blessing

- All the procedures below are justified using asymptotic theory for large n by assuming that $n \to \infty$, usually at some rate relative to T . Often n^2/T is treated as large in the asymptotics; this makes sense in an application with $T = 160$ and $n = 100$, say.

- By having large n, procedures (more sophisticated than the simple average in the previous example) are available for consistent estimation of tuning priors (prior hyperparameters) in forecasting and for factors in DFMs.

- Most of the theory, and all of the empirical work, has been developed within the past 10-12 years.

Outline

1) Why Might You Want To Use Hundreds of Series?
2) Dimensionality: From Curse to Blessing
3) **Dynamic Factor Models: Specification and Estimation**
4) Other High-Dimensional Forecasting Methods
5) Empirical Performance of High-Dimensional Methods
6) SVARs with Factors: FAVAR
7) Factors as Instruments
8) DSGEs and Factor Models

3) Dynamic Factor Models: Specification and Estimation

(A) Specification: The DFM, the Static Form, and the Approximate DFM

The idea (conjecture) behind DFMs is that small number of factors captures the covariation in macro time series (Geweke (1977), Sargent and Sims (1977)).

The exact DFM $X_{it} = \lambda_i(\mathrm{L})f_t + e_{it}$, $i = 1,\ldots,n$,

$\Psi(\mathrm{L})f_t = \eta_t$,

where:

$f_t = q$ unobserved "dynamic factors"

$\lambda_i(\mathrm{L})f_t$ = "common component"

$\lambda_i(\mathrm{L})$ = "dynamic factor loadings" lag polynomial

e_{it} = idiosyncratic disturbance

$\mathrm{cov}(f_t, e_{is}) = 0$ for all i, s

$Ee_{it}e_{js} = 0$, $i \neq j$, for all t, s (exact DFM)

The exact DFM, ctd.

DFM in vector notation: $\underset{n\times 1}{X_t} = \underset{n\times q}{\lambda(L)}\,\underset{q\times 1}{f_t} + \underset{n\times 1}{e_t}$

Identification of the factors: λ(L) and f_t are only identified up to a normalization: $\lambda(\mathrm{L})f_t = \lambda(\mathrm{L})HH^{-1}f_t$ for any square matrix H. This is unimportant if you are only interested in the space spanned by the *f*'s but it will come up in our discussion of FAVAR.

Spectral factorization. Because f_t and e_t are uncorrelated at all leads and lags, the spectrum of X_t is the sum of two components, one from the factors and one from the *e*'s:

$$S_{XX}(\omega) = \lambda(e^{i\omega})S_{ff}(\omega)\lambda(e^{-i\omega})' + S_{ee}(\omega),$$

where $S_{ee}(\omega)$ is diagonal under the exact DFM. This is the counterpart to the sum-of-variances expression in the cross-sectional factor model.

Forecasting in the exact DFM:

Consider forecasting X_{it+1} using all the data in X_t, and treat f_t as observed.

If u_{it} follows an autoregression and the errors are Gaussian, then

$$
\begin{aligned}
&E[X_{it+1}|\, X_t, f_t, X_{t-1}, f_{t-1}, \ldots] \\
&\quad = E[\lambda_i(\mathrm{L})f_{t+1} + e_{it+1}|\, X_t, f_t, X_{t-1}, f_{t-1}, \ldots] \\
&\quad = E[\lambda_i(\mathrm{L})f_{t+1}|\, X_t, f_t, X_{t-1}, f_{t-1}, \ldots] + E[e_{it+1}|\, X_t, f_t, X_{t-1}, f_{t-1}, \ldots] \\
&\quad = E[\lambda_i(\mathrm{L})f_{t+1}|\, f_t, f_{t-1}, \ldots] + E[e_{it+1}|\, X_t, f_t, X_{t-1}, f_{t-1}, \ldots] \\
&\quad = \alpha(\mathrm{L})f_{t+1} + \delta(\mathrm{L})X_{it} \qquad (1)
\end{aligned}
$$

- The *f*'s are sufficient for forecasting the *u*'s – they contain all the relevant information from the other *X*'s.
- The dimension reduction is from np parameters, to $(q+1)p$, where p is the number of lags.
- Under the DFM, the OLS dimension problem is eliminated and the forecast using the *f*'s will be first order efficient.

The approximate DFM

Chamberlain-Rothschild (1983)

Forni, Hallin, Lippi, Reichlin (2000, 2003a,b, 2004)

Stock and Watson (1999, 2002a,b)

The approximate DFM relaxes the strong assumption that the idiosyncratic terms are uncorrelated across equations at all leads and lags. The basic idea is that, instead of $S_{uu}(\omega)$ being diagonal, its eigenvalues are bounded as n increases (there is no linear combination that has increasing variance as n increases). Technical conditions will be displayed below when we go over asymptotics for DFMs.

The Static Form of the DFM

The DFM $X_t = \lambda(\mathrm{L})f_t + e_t$

where $\Psi(\mathrm{L})f_t = \eta_t$,

Suppose that $\lambda(\mathrm{L})$ has at most p_f lags. Then the DFM can be written,

$$\begin{pmatrix} X_{1t} \\ \vdots \\ X_{nt} \end{pmatrix} = \begin{pmatrix} \lambda_{10} & \cdots & \lambda_{1p_f} \\ \vdots & \ddots & \vdots \\ \lambda_{n0} & \cdots & \lambda_{np_f} \end{pmatrix} \begin{pmatrix} f_t \\ \vdots \\ f_{t-p_f} \end{pmatrix} + \begin{pmatrix} e_{1t} \\ \vdots \\ e_{nt} \end{pmatrix}$$

or $\underset{n\times 1}{X_t} = \underset{n\times r}{\Lambda} \underset{r\times 1}{F_t} + \underset{n\times 1}{e_t}$

where the number of static factors, r, could be as much as qp_f.

F_t is the vector of ***static factors***. The VAR for f_t implies that there is a VAR for F_t:

$$\Phi(\mathrm{L})F_t = G\eta_t$$

where G is a matrix of 1's and zeros and Φ consists of 1's, 0's, and Ψ's.

(B) Estimation: MLE, Principal Components, and Generalized PC

MLE Engle-Watson (1981); Stock and Watson (1989), Sargent (1989)

Suppose F_t follows a VAR(1). The DFM in static form is:

$$F_t = \Phi F_{t-1} + G\eta_t \qquad \text{(VAR(1) assumption)}$$
$$X_t = \Lambda F_t + e_t$$

Suppose that e_{it} follow individual AR's, written in first order form:

$$\tilde{e}_t = D\tilde{e}_{t-1} + H\zeta_t$$

where ζ_t is $n \times 1$, $H = [I_n \mid 0]'$, p_e is the number of lags in the e_{it} AR's, and $\tilde{e}_t = (e_t', e_{t\text{-}1}', \ldots, e_{t-p_e+1}')'$. Combining the F_t and $\tilde{e}_t$ equations yields:

MLE, ctd.

The DFM in state space form:

$$\begin{pmatrix} F_t \\ \tilde{e}_t \end{pmatrix} = \begin{pmatrix} \Phi & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} F_{t-1} \\ \tilde{e}_{t-1} \end{pmatrix} + \begin{pmatrix} G & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix} \tag{2}$$

$$X_t = \begin{bmatrix} \Lambda & \begin{pmatrix} I_n & 0 & \cdots & 0 \end{pmatrix} \end{bmatrix} \begin{pmatrix} F_t \\ \tilde{e}_t \end{pmatrix} \tag{3}$$

- Equation (2) is the state transition equation and equation (3) is the observer equation in the state space formulation of the DFM. The quasi-likelihood can now be computed using the Kalman filter.
- Early implementations used the MLE to estimate models with a single dynamic factor (r=1) with only a handful of variables:
  Engle-Watson (1981)
  Sargent (1989): estimate early DSGE
  Stock-Watson (1989): coincident index
  Quah-Sargent (1993) – more variables but a special structure

MLE, ctd.

- Historically, computation got too hard as *n* increased beyond a half-dozen variables (and the model was kept general), so other (nonparametric) methods were developed.
- However, there have been recent advances that make the MLE more practical:
  1) Computation
     a) faster computers
     b) can get very good starting values (specifics discussed next)
     c) new KF speedup: Jungbacker and Koopman (2008)
  2) Theory:
     Doz, Giannone, and Reichlin (2006)
  3) Empirical experience (discussed below):
     Doz, Giannone, and Reichlin (2006)
     Reiss and Watson (2008)

MLE, ctd.

The SS formulation of the DFM is particularly well suited to real time implementation – issues of irregular data arrival and mixed periodicity.

- The general setup for linear SS models with irregular data arrival is laid out in Harvey (1993).
- Aruoba, Diebold, and Scotti (2008) implement an irregular data/mixed frequency DFM by formulating the latent factor evolution at the daily level and the factors are observed either as temporal aggregates or at a point in time depending on the variable. In principle this system can provide internally consistent daily updates of economic conditions – indeed outlooks on hundreds of variables – along with forecast intervals with each new data release.

Estimation by Principal Components

DFM in static form: $X_t = \Lambda F_t + e_t$

By analogy to regression, consider estimating Λ and $\{F_t\}$ by NLLS:

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t)$$

subject to $\Lambda'\Lambda = I_r$ (identification). Now concentrate out $\{F_t\}$, given Λ:

$$\min_{\Lambda} T^{-1}\sum_{t=1}^{T} X_t'[I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda]X_t$$

$$\Leftrightarrow \max_{\Lambda} T^{-1}\sum_{t=1}^{T} X_t'\Lambda(\Lambda'\Lambda)^{-1}\Lambda X_t$$

$$\Leftrightarrow \max_{\Lambda} \mathrm{tr}\{(\Lambda'\Lambda)^{-1/2\prime}\,\Lambda'\left(T^{-1}\sum_{t=1}^{T} X_t X_t'\right)\Lambda(\Lambda'\Lambda)^{-1/2}\}$$

$$\Leftrightarrow \max_{\Lambda} \Lambda'\hat{\Sigma}_{XX}\Lambda \text{ s.t. } \Lambda'\Lambda = I_r, \text{ where } \hat{\Sigma}_{XX} = T^{-1}\sum_{t=1}^{T} X_t X_t'$$

$\Rightarrow \hat{\Lambda}$ = first r eigenvectors of $\hat{\Sigma}_{XX}$

$\Rightarrow \hat{F}_t = \hat{\Lambda}'X_t$ = first r principal components of X_t.

Distribution theory for PC as factor estimator

Results for the exact static factor model:

Connor and Korajczyk (1986)

- consistency in the exact static FM with T fixed, $n \to \infty$

Selected results for the approximate DFM: $X_t = \Lambda F_t + e_t$

Typical conditions (Stock-Watson (2002), Bai-Ng (2002, 2006),…):

(a) $\frac{1}{T}\sum_{i=1}^{T} F_t F_t' \xrightarrow{p} \Sigma_F$ (stationary factors)

(b) $\Lambda'\Lambda/n \to$ (or $\xrightarrow{p}$) Σ_Λ Full rank factor loadings

(c) e_{it} are weakly dependent over time and across series (approximate DFM)

(d) F, e are uncorrelated at all leads and lags

plus $n, T \to \infty$, with a relative rate condition

Selected results for the approximate DFM, ctd.

Stock and Watson (2002a)

- consistency in the approximate DFM, n, $T \to \infty$, no n/T restrictions
- justify using $\hat{F}_t$ as a regressor without adjustment

Bai and Ng (2006)

- $N^2/T \to \infty$ *(Think about this – not the principle of parsimony!)*
- asymptotic normality of PCA estimator of the common component at rate $\min(n^{1/2}, T^{1/2})$ in approximate DFM
- improve upon Stock-Watson (2002a) rate for using $\hat{F}_t$ as a regressor
- Method for constructing confidence bands for predicted value (these are for predicted value – *not* forecast confidence bands)

PC estimation in the approximate DFM, ctd.

- Data irregularities probably are best handled parametrically in the SS setup using the KF
- However the PC algorithm can be modified for data irregularities including mixed frequency data, see Stock and Watson (2002b, Appendix).

Generalized principal components

PC is motivated by considering a least squares problem. Presumably, if there is heteroskedasticity (or cross-correlation), you could do better by using WLS (or GLS) – which is what generalized PC does.

DFM in static form: $X_t = \Lambda F_t + e_t$

Infeasible WLS: Let Σ_{ee} be the variance matrix of e_t. The infeasible WLS estimator of F and W solves,

$$\min_{F_1,\ldots,F_T,\Lambda} \sum_{t=1}^{T} (X_t - \Lambda F_t)' \Sigma_{ee}^{-1} (X_t - \Lambda F_t).$$

Solution: $\hat{\Lambda}$ = first r eigenvectors of $\Sigma_{ee}^{-1/2} \hat{\Sigma}_{XX} \Sigma_{ee}^{-1/2\prime}$

and $\hat{F}_t = \hat{\Lambda}' X_t$ = first r generalized principal components of X_t.

Generalized principal components

Infeasible Generalized PC: $\hat{\Lambda}$ = first *r* eigenvectors of $\Sigma_{ee}^{-1/2}\hat{\Sigma}_{XX}\Sigma_{ee}^{-1/2\prime}$

Feasible Generalized PC requires an estimator of Σ_{ee}:

(a) Forni, Hallin, Lippi, and Reichlin (2005):

$$\hat{\Sigma}_{ee} = \hat{\Sigma}_{XX} - \hat{\Sigma}_{cc},$$

where $\hat{\Sigma}_{cc}$ is estimate of covariance matrix of the common component in the DFM, estimated by dynamic PCA (discussed below)

(b) Bovin and Ng (2003): $\hat{\Sigma}_{ee}^{diag} = \text{diag}(\hat{\Sigma}_{ee})$

(this accords from exact DFM restrictions)

(c) Stock and Watson (2005) – essentially "GLS by Cochrane-Orcutt"

Comments:

1. ***The basic idea – using factors as predictors***. Suppose the object is to forecast X_{it} using estimated factors. According to the exact DFM theory, the (first order) optimal forecast is obtained from the regression in (1). The dynamic factors aren't observed, so this leads to the regression,

$$X_{it+1} = \alpha(\mathrm{L})\hat{F}_t + \delta(\mathrm{L})X_{it} + \zeta_{t+1}$$

In some cases you might think some other variables W_t are good predictors so you could augment this:

$$X_{it+1} = \alpha(\mathrm{L})\hat{F}_t + \delta(\mathrm{L})X_{it} + \gamma(\mathrm{L})W_t + \zeta_{t+1}$$

If the number of regressors is small, this will yield first-order optimal forecasts.

Forecasting with estimated factors, ctd.

2. ***Multiple horizon forecasts.*** Two choices for *h*-step ahead forecasting:

   a. Direct forecasts:

$$X_{it+h} = \alpha(\mathrm{L})\hat{F}_t + \delta(\mathrm{L})X_{it} + \zeta^h_{t+h}$$

   b. Iterated forecasts:

$$X_{itt+1} = \alpha(\mathrm{L})\hat{F}_t + \delta(\mathrm{L})X_{it} + \zeta_{t+1}$$

$$\Phi(\mathrm{L})\hat{F}_{t+1} = \omega_{t+1} \quad (\text{VAR for } \hat{F}_{t+1}, \text{ where } \omega_{t+1} = G\eta_{t+1})$$

   Alternatively, the iterated forecasts can be implemented in the SS setup using the KF. The advantages and disadvantages of iterated v. direct were discussed in Lecture 10.

3. ***Forecast evaluation***: by simulated out of sample methods.

Dynamic Principal Components

This is something quite different – a way to extract principal components in the frequency domain (Brillinger (1964), discussed in Brillinger (1981))

- Dynamic PCA = PCA by frequency; the inverse Fourier Transform yields the dynamic principal components
- Two-sided projections yield common components
- *Distribution theory*:
  - Brillinger (1981) (asy. normality, n fixed, $T \to \infty$)
  - Forni, Hallin, Lippi, and Reichlin (2000) (consistency, n, $T \to \infty$)
  - Forni, Hallin, Lippi, and Reichlin (2004) (rates – optimal is $n \sim \sqrt{T}$ - slower rate because of estimation of the spectral density)
- DPCA pros and cons:
  - *pro:* nonparametric – no lag length restrictions needed
  - *con:* 2-sided $\hat{F}$*'s*, so not usable for second-stage regression

Which estimator to use – MLE, PC, or Generalized PC?

(a) Theoretical results ranking MLE, PC, and Generalized PC

Choi (2007) compares asymptotic variances of PC (derived by Bai (2003)) and Generalized PC, using the full covariance matrix of $e_t|(F_1,\ldots,F_T)$ (GLS, not WLS). Choi finds asymptotic gains for GPC (smaller variance of the asymptotic distribution for infeasible GPC than PC)).

Given the parameters, the KF estimator of F_t is the optimal estimator of F_t if the errors are Gaussian; for nonGaussian errors, the KF estimator is the MMSE estimator. This doesn't take parameter estimation error into account.

(b) Simulation evidence

- *Choi (2007)* compares PC, infeasible GLS-GPC, and feasible GLS-GPC in a MC study. He finds efficiency gains for feasible GPC in some cases, however the estimation of Σ hurts performance relative to infeasible GLS (Σ known), so feasible GPC improves on PC in some but not all cases. No evidence on full MLE.
- *Doz, Giannone, and Reichlin (2006)* MC study of:
  - PC
  - PC, estimation of DFM parameters using PC estimates, then a single pass of the Kalman Filter (Giannone, Reichlin, and Sala (2004))
  - ML (PC for starting values, then use EM algorithm to convergence)

Doz, Giannone, and Reichlin (2006) results for $\dfrac{tr\left(F'\hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'F\right)}{tr\left(F'F\right)}$

Table 2: Simulation results for the model: $\rho = .9$, $d = .5$, $\tau = .5$, $u = .1$, $r = 3$

	TR_{ml}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.48	0.59	0.65	0.67
$T = 100$	0.58	0.75	0.80	0.82
	Number of iterations			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	26	12	7	5
$T = 100$	20	9	5	4
	Computation time: seconds			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.72	0.46	0.56	1.44
$T = 100$	1.08	0.68	0.87	2.31
	TR_{ml}/TR_{pc}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.08	1.05	1.03	1.01
$T = 100$	1.10	1.06	1.02	1.01
	TR_{ml}/TR_{2s}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.05	1.02	1.01	1.00
$T = 100$	1.07	1.03	1.00	1.00

Table 3: Simulation results for the model: $\rho = .9$, $d = 0$, $\tau = 0$, $u = .1$, $r = 3$

	TR_{ml}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.54	0.65	0.68	0.70
$T = 100$	0.66	0.78	0.81	0.82
	Number of iterations			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	21	9	6	5
$T = 100$	15	7	5	4
	Computation time: seconds			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	0.58	0.36	0.49	1.30
$T = 100$	0.83	0.54	0.84	2.29
	TR_{ml}/TR_{pc}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.14	1.06	1.03	1.01
$T = 100$	1.19	1.06	1.02	1.01
	TR_{ml}/TR_{2s}			
	$n = 10$	$n = 25$	$n = 50$	$n = 100$
$T = 50$	1.07	1.02	1.01	1.00
$T = 100$	1.10	1.01	1.00	1.00

(b) Simulation evidence, ctd

- Boivin-Ng (2005) compare combinations of factor estimation methods and forecasting equation specifications, from the perspective of forecast MSE.
  - Of interest here is their comparison of PC (S, for static) to GPC (D, for dynamic)
  - The design in the following figures was calibrated to a large US macro data set
  - They report RMSE ratios, relative to an AR benchmark the columns to compare are the “S” (PC) to “D” (GPC) columns
  - Their conclusion is that PC generally works best.

Table 2a. RMSE for Calibrated DGP from 300 Simulations, Real Variables

Horizon	Variable	Forecasting Methods				
		SU/AR	SD/AR	SS/AR	DN/AR	DU/AR
1	IP	0.87	0.80	0.80	0.88	0.84
	GMYXSPQ	0.75	0.79	0.79	0.84	0.74
	MSMTQ	0.83	0.87	0.87	0.87	0.83
	LPNAG	0.79	0.71	0.71	0.84	0.79
2	IP	0.73	0.70	0.69	0.81	0.72
	GMYXSPQ	0.72	0.76	0.76	0.83	0.70
	MSMTQ	0.77	0.80	0.80	0.83	0.75
	LPNAG	0.83	0.70	0.69	0.86	0.83
4	IP	0.64	0.58	0.64	0.86	0.67
	GMYXSPQ	0.69	0.75	0.75	0.90	0.69
	MSMTQ	0.69	0.74	0.75	0.85	0.70
	LPNAG	0.67	0.63	0.65	0.74	0.69
6	IP	0.56	0.58	0.59	0.80	0.62
	GMYXSPQ	0.63	0.71	0.73	0.83	0.60
	MSMTQ	0.64	0.70	0.73	0.81	0.64
	LPNAG	0.57	0.69	0.65	0.69	0.59
8	IP	0.63	0.68	0.69	0.81	0.68
	GMYXSPQ	0.67	0.76	0.80	0.82	0.60
	MSMTQ	0.64	0.70	0.77	0.81	0.64
	LPNAG	0.64	0.81	0.71	0.69	0.62
12	IP	0.65	0.77	0.75	0.80	0.68
	GMYXSPQ	0.64	0.77	0.85	0.81	0.61
	MSMTQ	0.63	0.71	0.79	0.81	0.66
	LPNAG	0.61	0.88	0.76	0.73	0.61

Table 2b. RMSE for Calibrated DGP from 300 Simulations, Nominal Variables

Horizon	Variable	Forecasting Methods				
		SU/AR	SD/AR	SS/AR	DN/AR	DU/AR
1	PUNEW	0.91	0.93	0.93	0.93	0.91
	GMDC	0.88	0.90	0.90	0.90	0.89
	PUXX	0.97	0.98	0.98	0.97	0.97
	PWFSA	0.94	0.95	0.95	0.95	0.95
2	PUNEW	0.91	0.92	0.92	0.95	0.92
	GMDC	0.87	0.93	0.93	0.92	0.89
	PUXX	0.96	0.97	0.97	0.97	0.97
	PWFSA	0.94	0.95	0.95	0.96	0.94
4	PUNEW	0.84	0.88	0.88	0.92	0.86
	GMDC	0.78	0.94	0.94	0.88	0.80
	PUXX	0.95	0.97	0.97	0.96	0.96
	PWFSA	0.91	0.93	0.93	0.94	0.92
6	PUNEW	0.83	0.89	0.88	0.91	0.85
	GMDC	0.76	0.97	0.96	0.85	0.78
	PUXX	0.94	0.99	0.98	0.96	0.96
	PWFSA	0.91	0.94	0.94	0.94	0.92
8	PUNEW	0.86	0.93	0.91	0.92	0.88
	GMDC	0.77	1.03	0.99	0.85	0.80
	PUXX	0.95	1.03	1.01	0.95	0.96
	PWFSA	0.94	0.97	0.96	0.95	0.94
12	PUNEW	0.87	1.04	0.96	0.92	0.90
	GMDC	0.76	1.13	1.04	0.83	0.83
	PUXX	0.94	1.10	1.04	0.95	0.97
	PWFSA	0.94	1.02	0.97	0.95	0.95

(c) Empirical evidence

(i) Comparisons of forecasts – actual data sets (US, EU):

- Stock & Watson, *Handbook of Economic Forecasting* (2006) plus extensive empirical work as backup – empirical forecasting comparison over many series
- D'Agostino and Giannone (2006)
- Marcellino and coauthors (several)
- Broad summary of findings across papers:
  - PC, WLS-PC, and GLS-PC have very similar performance.
  - GLS-PC can produce modest outliers (sometimes better, sometimes worse)
  - mild preference for PC

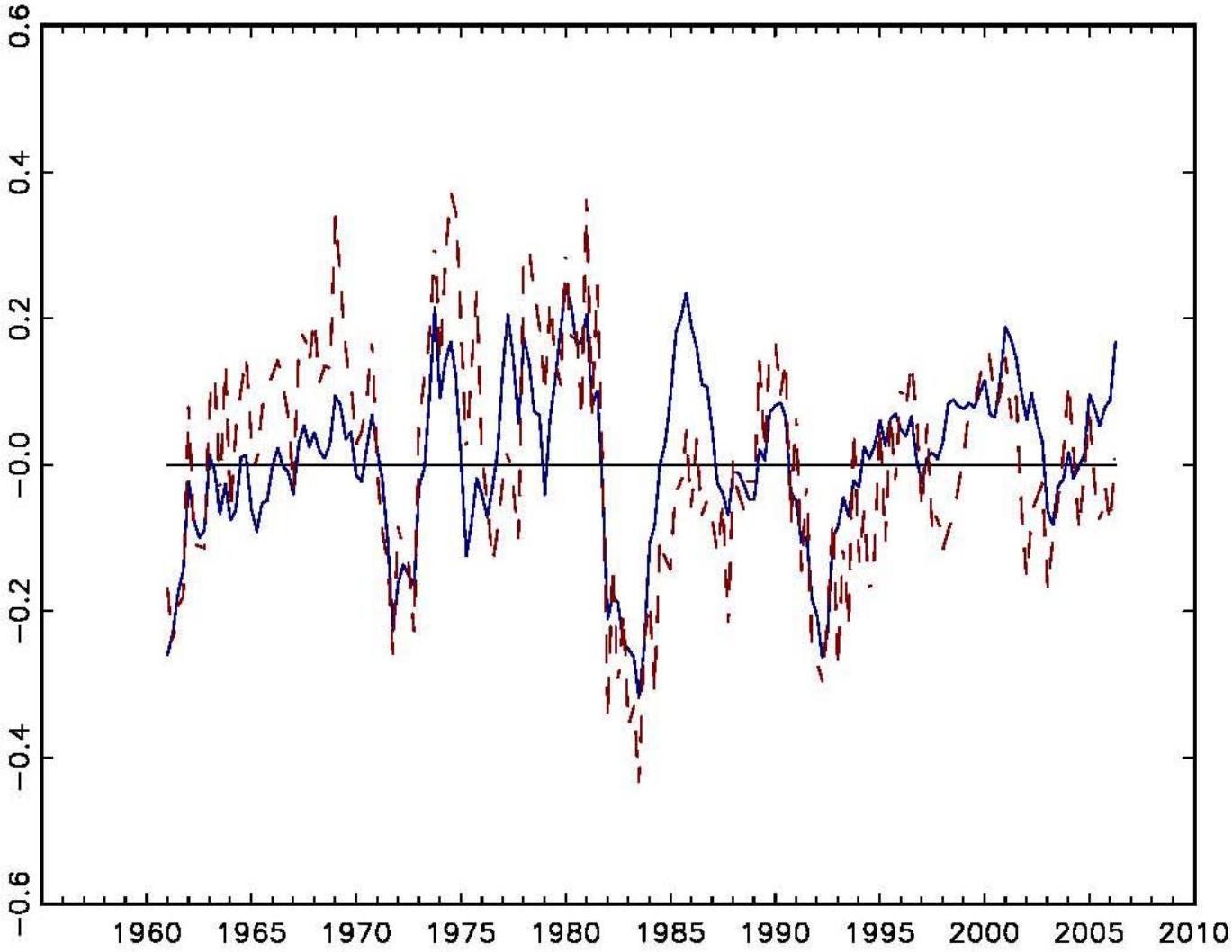
(c) Empirical evidence, ctd

(ii) A bit of filtering evidence

- Riess and Watson (2007)
  - application in which the factor structure is weak (prices with large idiosyncratic terms – lots of idiosyncratic movement + measurement error)
  - PC estimate of factors vs. MLE from the KF – take a look!

Figure 4. Alternative estimates of pure inflation ($v_t - v_{t-8}$):

Benchmark parametric model (solid blue) and principal components (dashed red)



Which estimator to use?

- For forecasting, it doesn't seem to matter much – PC seems to work as well as the others in typical applications

- MLE is appealing theoretically and has the additional advantage of temporal smoothing – this might be the most promising avenue currently.

Selecting the number of factors

DFM in static form: $X_t = \Lambda \overset{r\times 1}{F_t} + e_t$

What is r?

Will discuss:

1) Informal data analysis
2) Estimating the number of static factors
   a. Estimation of r
   b. Testing $r = r_0$ v. $r > r_0$
3) Estimating the number of dynamic factors, q

(1) Informal data analysis

- Largest eigenvalues
- scree plots (plots of ordered eigenvalues of $X'X/T$
- fraction of trace R^2 explained

(2) Estimating the number of static factors

Estimation approach

Bai-Ng (2002) propose an estimator of r based on an information criterion; their main result is $\hat{r} \xrightarrow{p} r_0$ for the approximate DFM

Digression on information criteria (IC) for lag length selection in an AR

Consider the AR(p): $y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + \varepsilon_t$

- Why not just maximize the R^2?
- IC trades off estimator bias (too few lags) vs. estimator variance (too many lags), from the perspective of fit of the regression:

Bayes Information Criterion: $\mathrm{BIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{\ln T}{T}$

Akaike Information Criterion: $\mathrm{AIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{2}{T}$

The Bayes Information Criterion (BIC)

$$\text{BIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{\ln T}{T}$$

- *First term*: always decreasing in p (larger p, better fit)
- *Second term*: always increasing in p.
  - The variance of the forecast due to estimation error increases with p
  - This term is a "penalty" for using more parameters
  - The penalty gets smaller with the sample size
- *Minimizing BIC*(p) trades off bias and variance to determine a "best" value of p for your forecast.
  - The result is that $\hat{p}^{BIC} \xrightarrow{p} p$
  - In theory, any penalty $g(T) \to 0$, $Tg(T) \to \infty$ will produce $\hat{p} \xrightarrow{p} p_0$
  - Method of proof: show (i) $\Pr[\hat{p}^{BIC} < p] \to 0$; (ii) $\Pr[\hat{p}^{BIC} > p] \to 0$ (proof in (*SW*, *Introduction to Econometrics*, *App. 14.5*))

The Akaike Information Criterion (AIC)

$$\text{AIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{2}{T}$$

$$\text{BIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + p\frac{\ln T}{T}$$

The penalty term is smaller for *AIC* than *BIC* ($2 < \ln T$)

- *AIC* estimates more lags (larger p) than the *BIC*
- In fact, the AIC estimator of p isn't consistent – it can overestimate p – the penalty isn't big enough: for AIC, $Tg(T) = T \times (2/T) = 2$, but you need $Tg(T) \to \infty$ for consistency.
- Still, AIC might be desirable if you want to err on the side of long lags

Example: AR model of U.S. Δinflation, lags 0 – 6:

# Lags	BIC	AIC	R^2
0	1.095	1.076	0.000
1	1.067	1.030	0.056
2	**0.955**	0.900	0.181
3	0.957	**0.884**	0.203
4	0.986	0.895	0.204
5	1.016	0.906	0.204
6	1.046	0.918	**0.204**

- BIC chooses 2 lags, AIC chooses 3 lags.
- If you used the R^2 to enough digits, you would (always) select the largest possible number of lags.

Estimating the number of static factors, ctd.

The Bai-Ng (2002) information criteria have the same form:

$$\text{IC}(r) = \ln\left(\frac{SSR(r)}{T}\right) + \text{penalty}(N, T, r)$$

Bai-Ng (2002) propose several IC's with different penalty factors that all produce consistent estimators of k. Here is the one that seems to work best in MCs (and is the most widely used in empirical work):

$$IC_{p2}(r) = \ln(\text{V}(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where $$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

$\hat{F}_t^r$ are the PC estimates of r factors

(*minor notational note*: Bai-Ng (2002) use proxy argument k, not r)

Estimating the number of static factors, ctd.

Bai-Ng (2002) IC_{p2}: $$IC_{p2}(r) = \ln(V(r,\hat{F}^r)) + r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right]$$

where $$V(r,\hat{F}^r) = \min_{\Lambda}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it} - \lambda_i^{r\prime}\hat{F}_t^r\right)^2$$

Comments:

- $\ln(V(r,\hat{F}^r))$ is a measure of (trace) fit – generalizes $\ln(SSR/T)$ in the BIC
- If $N = T$, then $r\left(\frac{N+T}{NT}\right)\ln\left[\min(N,T)\right] = r\left(\frac{2T}{T^2}\right)\ln T = 2r\frac{\ln T}{T}$

  which is 2 × the usual BIC penalty factor
- Both N and T are in the penalty factor: you need $N, T \to \infty$.
- Bai-Ng's (2002) main result: $\hat{r} \xrightarrow{p} r_0$
- Logic of proof is same as for BIC
- In practice, different IC can yield substantially different answers

(3) Estimating the number of dynamic factors, q

Bai-Ng consider estimating the number of static factors – which is directly useful for forecasting using PC.

For the MLE (which specifies a process for the dynamic factors) it is desirable to estimate the number of dynamic factors. Recall that the static factors are constructed by stacking the dynamic factors:

$$F_t = \begin{pmatrix} f_t \\ \vdots \\ f_{t-p_f} \end{pmatrix}$$

so the static factors must be dynamically singular: the rank of the innovation variance matrix in the projection of F_t on F_{t-1} must be the rank of (the spectrum of) f_t (since many of the elements of F_t are perfectly predictable from F_{t-1})

Estimating the number of dynamic factors, ctd:

Three ways to test for this dynamic singularity:

- Amenguel-Watson (2007)

  Regress X_t on $\hat{F}_{t-1}$; the residuals will have factors of rank of the dynamic factors, use Bai-Ng (2002) to estimate that rank

- Bai and Ng (2007)

  Estimate a VAR for $\hat{F}_t$, then estimate the rank of the residual covariance matrix

- Hallin and Liška (2007)

  Frequency domain (rank of spectrum of X_t will be number of dynamic factors)

Testing approach to determining k

- This is a very difficult problem!

- Consider testing $k = 0$ v. $k > 0$. If $k = 0$ then the $n \times n$ variance matrix of X_t has no dominant eigenvalues. Thus testing $k = 0$ v. $k > 1$ entails comparing the largest eigenvalue of $X'X/T$ (where each X_i has been standardized) to a critical value.

- The exact finite sample theory in the i.i.d. standard normal case is based on eigenvalues of Wishart distributions (see Anderson (1984). That distribution (i) hinges on normality and (ii) is sensitive to misspecification of the variance matrix of X.

Testing approach, ctd.

- Work in this area has focused on generalizing/extending this to large random matrices
  - Tracy-Widom (1994): distribution of largest eigenvalue of $X'X/T$, X_{it} i.i.d. N(0,1)
  - Johnstone (2001), El Karoui (2007): Tracy-Widom for largest eigenvalue under weaker assumptions
  - Onatski (2007): joint Tracy-Widom for m largest eigenvalues under weaker assumptions (distribution of scree plot)
  - Onatski (2008): testing H_0: $k = k_0$ v. $k > k0$ in DFM

- This research program is incomplete, but it holds the promise of (some day) providing a more refined method for determining k than IC