AEA Continuing Education Course

Time Series Econometrics

Lectures 8 and 9

James H. Stock

January 7, 2010

Forecasting and Macro Modeling with Many Predictors: Introduction to Dynamic Factor Models

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

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

One of the major challenges of empirical macro is that there is limited information – limited historical experience. But thousands of economic time series are available on line in real time. Can these be used to expand our information 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.

Four 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.

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

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

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.
- OLS is a bad idea with many regressors: in theory, if regressors are proportional to sample size, consistency is lost; in practice, the problem is introducing large estimation error.

The blessing part (one view)

In some models – dynamic factor models in particular – many series *helps* to identify the statistical object of interest – inference can be *improved* when there are many series. (Geweke's (1993) discussion of Quah and Sargent (1993)).

From curse to blessing: dynamic factor models (Geweke (1977), Sargent and Sims (1977))

Suppose the n variables in X_t are related to q unobserved factors f_t, which evolve according to a time series process:

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

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

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.

An example following Forni and Reichlin (1998).

Suppose f_t is scalar and $\lambda_i(\mathrm{L}) = \lambda_i$ ("no lags in the factor loadings"), so

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

Then

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

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: dynamic factor models, ctd

- 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 $n = 130$, 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) SVARs with Factors: FAVAR
5) Factors as Instruments
6) DSGEs and Factor Models
7) Other High-Dimensional Forecasting Methods
8) Empirical Performance of High-Dimensional Methods

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 spectral density 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

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

- The f's 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 of u_t 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 ("little *f* and big *F*")

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: Principal Components, Generalized PC, and MLE

(i) Estimation by Principal Components

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

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

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

$$\min_{F_1,\ldots,F_T,\Lambda} T^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t) \qquad (2)$$

subject to $\Lambda'\Lambda = I_r$ (identification). Given Λ, the (infeasible) OLS estimator of F_t is:

$$\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$$

Now substitute $\hat{F}_t(\Lambda)$ into (2) to concentrate out $\{F_t\}$:

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

Principal components, ctd.

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

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

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

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

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

Remember $\hat{F}_t(\Lambda) = \left(\Lambda'\Lambda\right)^{-1}\Lambda' X_t$, so

$$\hat{F}_t(\hat{\Lambda}) = \left(\hat{\Lambda}'\hat{\Lambda}\right)^{-1}\hat{\Lambda}' X_t = \hat{\Lambda}' X_t \quad \text{(because } \hat{\Lambda}'\hat{\Lambda} = I_r\text{)}$$

= 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)

- $N^2/T \to \infty$ *(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).

(ii) 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, ctd.

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"

(iii) MLE

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

Suppose f_t follows a VAR(1). The DFM with first order dynamics is:

$$f_t = \Psi f_{t-1} + \eta_t \qquad \text{(VAR(1) assumption)}$$
$$X_t = \lambda_0 f_t + \lambda_1 f_{t-1} + e_t$$

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:

$$\begin{pmatrix} f_t \\ f_{t-1} \\ \tilde{e}_t \end{pmatrix} = \begin{pmatrix} \Psi & 0 & 0 \\ I & 0 & 0 \\ 0 & 0 & D \end{pmatrix} \begin{pmatrix} f_{t-1} \\ f_{t-2} \\ \tilde{e}_{t-1} \end{pmatrix} + \begin{pmatrix} I & 0 \\ 0 & 0 \\ 0 & H \end{pmatrix} \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix} \tag{3}$$

$$X_t = \begin{bmatrix} \lambda_0 & \lambda_1 & (I_n & 0 & \cdots & 0) \end{bmatrix} \begin{pmatrix} f_t \\ f_{t-1} \\ \tilde{e}_t \end{pmatrix} \tag{4}$$

- Equation (3) is the state transition equation and equation (4) is the observer equation in the state space formulation of the DFM. The quasi-likelihood can now be computed using the Kalman filter.
- Extension to higher order dynamics (higher order VAR for $\Psi(L)$ and/or lags of f_t entering X_t equation): augment state vector with further lags of f_t

MLE, ctd.

- 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

- 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.

MLE, ctd.

- However, there have been recent advances that make the MLE more practical:
  1) Computation
     a) faster computers
     b) can get very good starting values from PC: estimate factors $\hat{F}_t$, then estimate parameters treating $\hat{F}_t$ as observed data
     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.

Forecasting with estimated factors

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 are an empirical matter (see Marcellino, Stock, & Watson (2006), Pesaran, Pick, and Timmerman (2009)).

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

Digression: 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 $\frac{tr\left(F'\hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'F\right)}{tr\left(F'F\right)}$

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: SU is PC using unrestricted forecasting equation, DU is GPC using unrestricted forecasting equation
  - 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):

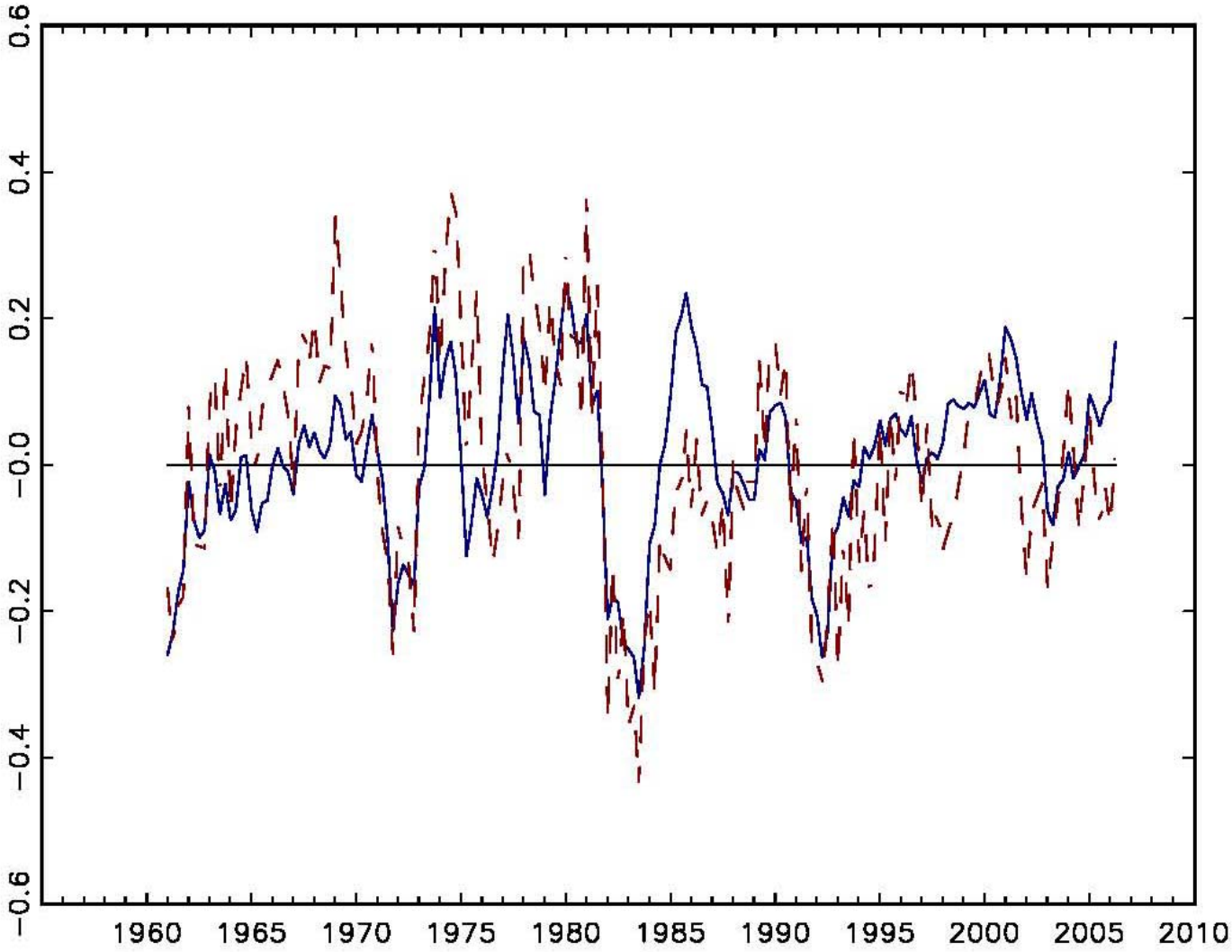
- Stock & Watson, *Handbook of Economic Forecasting* (2006a) plus extensive empirical work as backup – empirical forecasting comparison over many series
- D'Agostino and Giannone (2006)
- Marcellino and coauthors (several)
- Broad summary of findings across papers:
  - PC, WLS-PC, and GLS-PC have very 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 seems to 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

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

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

$\hat{F}_t^r$ 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

Comments on Bai-Ng factor selection

- Monte Carlo studies show B-N works well when n, T are large
- But in practice:
  - Different IC can yield substantially different answers
  - Adding series often increases the number of estimated factors (adding sectors should increase number of factors; adding series within sectors should not)
- Various possibilities:
  - More persistence in the data than in simulations (highly persistent idiosyncratic terms) – Greenaway-McGrevy, Han, and Sul (2008) – maybe prefilter? (prefiltering discussed below)
  - Maybe there are many factors of moderate importance (not just a few dominant factors) – examine empirically (discussed below)
- Judgment is required

(3) Estimating the number of dynamic factors, q

Bai-Ng consider estimating the number of static factors (r) – 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 (q). 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 (as opposed to *IC* approach) to determining q

- This is a very difficult problem!

- Consider testing $q = 0$ v. $q > 0$. If $q = 0$ then the $n \times n$ variance matrix of X_t has no dominant eigenvalues. Thus testing $q = 0$ v. $q > 1$ 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: $r = r_0$ v. $r > r_0$ in DFM
  - Harding (2009): selection algorithm based on Tracy-Widom

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

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

4) SVARs with Factors: FAVAR

Challenges & critiques of standard SVAR modeling include:

- The Rudebush (1998) critique of SVARs with short-run timing identification: Fed uses more information than is in a standard VAR
- The invertibility problem in SVARs: is $Ru_t = \varepsilon_t$, $\varepsilon_t = R^{-1}u_t$ plausible?
- Including more variables in the VAR might improve forecast efficiency and provide an internally consistent set of forecasts for a large number of variables – but confronts the n^2p parameter problem

Bernanke, Boivin, and Eliasz's (2005) (BBE) idea is to use factors as a way to solve this problem: in a DFM, factors summarize all the relevant information on the economy. The result is the BBE Factor Augmented VAR (FAVAR).

FAVAR, ctd

There are a number of ways FAVAR can be implemented, the following papers use related approaches but differ in the details:
Bernanke, B.S., and J. Boivin (2003), Bernanke, Boivin, and Eliasz (2005) (BBE), Favero and Marcellino (2001), Favero, Marcellino, and Neglia (2004); also see Giannone, Reichlin, and Sala (2004) on the invertibility issue.

Here we follow the spirit of BBE (2005) although some technical details (but not identification ideas) are different – this development follows Stock and Watson (2005).

One approach would be simply to put factors into a SVAR, however the factors themselves are not identified so making any identification assumptions about their innovations is difficult.

FAVAR, ctd.

VAR form of the exact DFM

DFM with first order dynamics from above:

$$F_t = \Phi F_{t-1} + \mathrm{G}\eta_t$$

$$X_t = \Lambda F_t + e_t$$

$$e_t = De_{t-1} + \zeta_t$$

where D is diagonal. Quasi-difference X_t:

$$(I - D\mathrm{L})X_\mathrm{t} = (I - D\mathrm{L})\Lambda F_t + \zeta_t = \Lambda F_t - D\Lambda F_{t-1} + \zeta_t$$

Substitute in $F_t = \Phi F_{t-1} + \mathrm{G}\eta_t$:

$$(I - D\mathrm{L})X_\mathrm{t} = \Lambda(\Phi F_{t-1} + \mathrm{G}\eta_t) - D\Lambda F_{t-1} + \zeta_t$$

Rearrange:

$$X_\mathrm{t} = (\Lambda\Phi - D\Lambda)F_{t-1} + DX_{t-1} + \Lambda\mathrm{G}\eta_t + \zeta_t$$

Putting the F_t and X_t equations together yields,

VAR form of the DFM, ctd.

$$\begin{pmatrix} F_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Phi & 0 \\ \Lambda\Phi - D\Lambda & D \end{pmatrix} \begin{pmatrix} F_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} G & 0 \\ \Lambda G & I \end{pmatrix} \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix}$$

Writing the reduced form VAR as $A(L)X_t = u_t$, the VAR innovations are $u_t = X_t - \text{Proj}(u_t|F_{t-1}, F_{t-2},\ldots, X_{t-1}, X_{t-2},\ldots) = \Lambda G\eta_t + \zeta_t$, where we are treating the F's as observed (this is justified by large n asymptotics).

The ζ's are disturbances to the idiosyncratic process. What we are interested in is the response of X_t to structural shocks, which affect all the variables. The structural shocks ε_t are related to the innovations in the dynamic factors:

$$R\eta_t = \varepsilon_t \qquad \text{(structural model in SVAR)}$$

FAVAR

reduced form:
$$\begin{pmatrix} F_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Phi & 0 \\ \Lambda\Phi - D\Lambda & D \end{pmatrix} \begin{pmatrix} F_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} G & 0 \\ \Lambda G & I \end{pmatrix} \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix}$$

structure:
$$\underset{q\times q}{R}\ \underset{q\times 1}{\eta_t} = \underset{q\times 1}{\varepsilon_t}$$

The structural IRF is the distributed lag of X_t on ε_t. Now

$$X_t = \Lambda F_t + e_t$$

and
$$F_t = \Phi F_{t-1} + G\eta_t = \Phi F_{t-1} + GR^{-1}\varepsilon_t,$$

so
$$X_t = \Lambda(I - \Phi L)^{-1} GR^{-1}\varepsilon_t + e_t$$

so the structural IRF is $\Lambda(I - \Phi L)^{-1} GR^{-1}$.

FAVAR, ctd.

Comments:

1.***Lags***. These formulas are for first order dynamics – with higher order dynamics the expression above becomes,

$$\begin{pmatrix} F_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Phi(L) & 0 \\ \Lambda\Phi(L) - D(L)\Lambda & D(L) \end{pmatrix} \begin{pmatrix} F_{t-1} \\ X_{t-1} \end{pmatrix} + \begin{pmatrix} G & 0 \\ \Lambda G & I \end{pmatrix} \begin{pmatrix} \eta_t \\ \zeta_t \end{pmatrix}$$

structure: $\underset{q\times q}{R}\ \underset{q\times q}{\eta_t} = \underset{q\times q}{\varepsilon_t}$

2.***Identification***. The identification problem is finding R, where $R\eta_t = \varepsilon_t$. This is now amenable to applying the SVAR identification toolkit:

- Timing scheme (CEE (1999), BBE (2005): slow/policy/fast)
- long run restrictions (Blanchard-Quah (1989); Gali (1999)
- sign restrictions (Uhlig (2005), Ahmadi and Uhlig (2007))
- heteroskedasticity (Rigobon (2003), Rigobon-Sack (2003))

FAVAR, ctd.

3. ***Structural shocks.*** The η_t shocks are the shocks to the *dynamic* factors: $F_t = \Phi F_{t-1} + \mathrm{G}\eta_t$. These are *not* the residuals from a VAR estimated using F_t: the number of static factor innovations $r \geq q$. Implementation involves estimating the space of dynamic factor shocks, which in turn entails (i) estimating the number of dynamic factors q, and (ii) reduced rank regressions to estimate η_t.

4. ***Many impulse responses.*** The structural IRF is $\Lambda(I - \Phi L)^{-1} GR^{-1}$, which yields IRFs for all the X's in the system!

FAVAR, ctd.

5. ***Overidentification.*** These systems move from being exactly identified SVARs to potentially heavily overidentified. Consider the BBE fast/slow identification idea: the slow identification restriction now applies to a huge block of variables, specifically, ε_t^r should not load on any of the slow moving variables. Let u_t^S be the VAR innovations to the slow-moving variables, $u_t^S = X_t^S - \text{Proj}(X_t^S | F_{t-1}, F_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots)$. Under the fast/slow identification scheme, $\text{Proj}(u_t^S | \varepsilon_t^r)$ should be zero. These many overidentifying restrictions are testable.

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

5) Factors as Instruments

Independently developed by Kapetanios and Marcellino (Oct. 2006, revised 2008) and Bai and Ng (Oct. 2006, revised 2007b)

Digression on weak instruments: Weak instruments arise if the IV is weakly correlated with the included endogenous regressor.

- Using factors might be a way to use more information
- The instruments $\hat{F}_t$ are linear combinations of the X_t's, but the key insight is that the coefficients of that linear combination are estimated separately, not in the first-stage regression (the X's don't enter the moment conditions explicitly).
- The mathematics is essentially the same as the math used to show that $\hat{F}_t$ can be used in a forecasting regression without a generated regressor problem.

Factors as instruments, ctd.

Main result: under conditions like those above (the approximate DFM conditions), and the "usual" large-*n* rate condition $N^2/T \to \infty$, and a strong instrument assumption,

$$\sqrt{T}\,[\hat{\beta}^{TSLS}(F_t) - \hat{\beta}^{TSLS}(\hat{F}_t)] \xrightarrow{p} 0 \qquad (5)$$

where $\hat{F}_t$ is the PC estimator of the factors. So IV is as efficient if the factors are known as if they are not when *N* is large.

Simulation results in Kapetanios and Marcellino (2008) and Bai and Ng (2007b) are promising concerning the finite-sample validity of (5) under strong instruments.

Factors as instruments, ctd.

Additional comments

1. The idea of using principal components as instruments is old (Kloek and Mennes (1960), Amemiya (1966)) – what is new is proving optimality and distribution results using the DFM as the conceptual framework.
2. Not all the individual X's need to be valid instruments – the e's could be correlated with the included endogenous regressor, what matters is that the F's are not correlated.
3. If there isn't a factor structure, then the PC estimates are going to random linear combinations of the X's. But if the X's are all valid instruments, the $\hat{F}_t$'s remain valid instruments even without a factor structure (details in Bai and Ng (2007)).
4. If the instruments (F's) are weak, then weak instrument considerations kick in. (The original hope is that weak instruments will be less of a problem using the F's.)

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

6) DSGEs and Factor Models

"Reduced form" DFM with first order dynamics from above:

$$F_t = \Phi F_{t-1} + G\eta_t$$
$$X_t = \Lambda F_t + e_t$$
$$e_t = De_{t-1} + \zeta_t$$

Boivin and Giannoni (2006b) replace the reduced form state space model with a linearized DSGE:

$$\tilde{F}_t = \tilde{\Phi}\,\tilde{F}_{t-1} + \tilde{G}\,\tilde{\eta}_t \qquad (6)$$
$$X_t = \tilde{\Lambda}\,\tilde{F}_t + e_t, \qquad (7)$$
$$e_t = De_{t-1} + \zeta_t \qquad (8)$$

where ~ means that (6) is a structural model (DSGE), cf. Sargent (1989), Boivin-Giannoni (2006b).

DSGEs and factor models, ctd.

$$\tilde{F}_t = \tilde{\Phi}\,\tilde{F}_{t-1} + \tilde{G}\,\tilde{\eta}_t$$

$$X_t = \tilde{\Lambda}\,\tilde{F}_t + e_t,$$

$$e_t = De_{t-1} + \zeta_t$$

The DSGE implies restrictions on $\tilde{\Lambda}$ that identify $\tilde{F}_t$:

- The elements of $\tilde{F}_t$ are the state vector of the DSGE, for example, $\tilde{F}_t = (x_t,\ \pi_t,\ r_t,\ rr_t,\ \Delta a_t,\ u_t,\ \tau_t)'$.
- The meanings of the elements $\tilde{F}_t$ within the DSGE imply restrictions on $\tilde{\Lambda}$ that identify $\tilde{F}_t$
- The system, with restrictions on $\tilde{\Lambda}$ imposed, is in SS form and the KF can be used to compute the likelihood. Estimation is a combination of DFM MLE and DSGE MLE with a small number of variables:
  - initial values using PC estimates of the factors
  - modified Jungbacker-Koopman (2008) speedup?

Boivin-Giannoni (2006b) identification: Setup: let $\tilde{\lambda}$ denote a nonzero entry (not all the same – just dropping subscripts)

$$\begin{bmatrix} \text{output gap series \#1} \\ \vdots \\ \text{output gap series \#n}_{\text{Y}} \\ \text{inflation series \#1} \\ \vdots \\ \text{inflation series \#n}_{\text{C}} \\ \vdots \\ \text{----} \\ \text{Information series \#1} \\ \vdots \\ \text{Information series \#n}_{\text{info}} \end{bmatrix} = \begin{bmatrix} \tilde{\lambda} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\lambda} & 0 & \cdots & 0 \\ 0 & \tilde{\lambda} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \tilde{\lambda} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\lambda} & \tilde{\lambda} & \cdots & \tilde{\lambda} \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{\lambda} & \tilde{\lambda} & \cdots & \tilde{\lambda} \end{bmatrix} \begin{bmatrix} x_t \\ \pi_t \\ \vdots \\ \tilde{F}_{last,t} \end{bmatrix}, \text{ where } \tilde{F}_t = \begin{bmatrix} x_t \\ \pi_t \\ \vdots \\ \tilde{F}_{last,t} \end{bmatrix}$$

Or

$$\begin{bmatrix} X_{\text{sensor},t} \\ X_{\text{info},t} \end{bmatrix} = \begin{bmatrix} \tilde{\Lambda}_{\text{sensor}} \\ \tilde{\Lambda}_{\text{info}} \end{bmatrix} \tilde{F}_t, \text{ where } \tilde{F}_t = \tilde{\Phi}(\mathrm{L})\tilde{F}_t + \tilde{\varepsilon}_t t$$

In general the information series can have weights on expectations of future F_t (e.g. term spreads) but by the VAR structure of the factors plus the DFM assumptions those are projected back on F_t.

Results from Boivin-Giannone (2006b) (they use Bayes methods)

- Case A: 7 variables
- Case B: 14 variables
- Case C: 91 variables

Table 1: Priors and estimates of structural parameters

	Prior Distribution			SW	Case A	Case B	Case C
	Type	Mean	St.Err.				
φ	Normal	4	1.5	5.36	5.88	6.17	3.81
				(0.88)	(1.11)	(1.13)	(1.04)
σ_c	Normal	1	0.375	1.54	1.45	1.79	1.63
				(0.24)	(0.23)	(0.44)	(0.44)
h	Beta	0.7	0.1	0.71	0.75	0.54	0.50
				(0.07)	(0.07)	(0.27)	(0.27)
σ_L	Normal	2	0.75	2.34	2.18	2.42	2.41
				(0.60)	(0.65)	(0.69)	(0.68)
ϕ	Normal	1.25	0.125	1.42	1.24	1.37	1.26
				(0.08)	(0.07)	(0.07)	(0.07)
$1/\psi$	Normal	0.2	0.075	0.32	0.27	0.26	0.27
				(0.06)	(0.06)	(0.06)	(0.06)
ξ_ω	Beta	0.75	0.05	0.81	0.77	0.78	0.82
				(0.02)	(0.03)	(0.04)	(0.03)
ξ_p	Beta	0.75	0.05	0.88	0.90	0.88	0.86
				(0.01)	(0.02)	(0.01)	(0.02)
γ_ω	Beta	0.5	0.15	0.39	0.45	0.43	0.48
				(0.12)	(0.14)	(0.14)	(0.14)
γ_p	Beta	0.5	0.15	0.66	0.72	0.50	0.36

				(0.12)	(0.14)	(0.14)	(0.14)
γ_p	Beta	0.5	0.15	0.66	0.72	0.50	0.36
				(0.08)	(0.19)	(0.15)	(0.14)
ρ	Beta	0.75	0.1	0.76	0.67	0.72	0.70
				(0.02)	(0.05)	(0.04)	(0.03)
$r_{\pi 0}$	Normal	1.8	0.1	1.78	1.81	1.72	1.66
				(0.08)	(0.10)	(0.10)	(0.09)
$r_{\pi 1}$	Normal	-0.3	0.1	-0.22	-0.22	-0.30	-0.39
				(0.09)	(0.12)	(0.10)	(0.09)
r_{y0}	Normal	0.188	0.05	0.22	0.23	0.24	0.22
				(0.03)	(0.03)	(0.03)	(0.03)
r_{y1}	Normal	-0.063	0.05	-0.13	-0.11	-0.12	-0.12
				(0.03)	(0.03)	(0.04)	(0.03)
Implied parameters							
pseudo EIS: $\frac{1-h}{(1+h)\sigma_c}$				0.110	0.099	0.167	0.204
slope of PC: $\frac{(1-\beta\xi_p)(1-\xi_p)}{(1+\beta\gamma_p)\xi_p}$				0.011	0.007	0.012	0.018

The parameter estimates are given by the median of the posterior distribution
Results are based on 100 000 replications. Standard errors are reported in ().

Misc. concluding DFM comments

1. Everything in this lecture has applied to variables with short-run dependence. There is a fair amount of work extending DFMs to handle unit roots and cointegration, see Bai and Ng (2004) and Banerjee and Marcellino (2008, 2009).

2. We also have ignored TVP and structural breaks in DFMs. DFMs have a certain robustness to TVP and structural breaks, however the only published work with any TVP aspect in DFMs is Stock and Watson (2002) and Phillips and Sul (1997). Recent work includes Stock and Watson (2009) and Banerjee, Marcellino, and Masten (2007).

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

7) Other High-Dimensional Forecasting Methods

- High-dimensional prediction has received a great deal of attention in the past decade in the statistics literature and many new methods have been developed.

- ***Hybrid DFMs***. e.g. incorporating cointegrating relations into DFMs (DFMs meet ECMs), see Banderjee, Marcellino and Masten (2009)

- ***Other approaches***. Stock and Watson (2004) provide a partial survey. Here are some of the methods and some references:

  1. ***Bayesian VARs*** (strong priors over many parameters) De Mol, Giannone, and Reichlin (2008), Carriero, Kapetanios, and Marcellino (2009)

Non-DFM high-dimensional methods, ctd.

2. ***Bayesian model averaging (BMA).*** Leamer (1978); Min and Zellner (1990); Fernandez, Ley, and Steele (2001a,b), Koop and Potter (2004); Clyde, Desimone, and Parmigiani (1996), Clyde (1999). Surveys: Hoeting, Madiga, Raftery, and Volinsky (1999), Geweke and Whiteman (2004)

3. ***Empirical Bayes*:** Robbins (1964), Efron and Morris (1973), Edelman (1988), Zhang (2003, 2005); Maritz and Lwin (1989), Carlin and Louis (1996), and Lehmann and Casella (1998, Section 4.6)

4. ***Bagging***: Breiman (1996), Bühlmann and Yu (2002); Inoue and Kilian (2008)

Non-DFM high-dimensional methods, ctd.

5. ***LASSO & boosting*:** Econometric applications: Bai and Ng (2007), De Mol, Giannone, and Reichlin (2006), Bai and Ng (2007)

6. ***Hard thresholding/false discovery rate methods*:** information criteria through… for high level connections see Efron (2003)

7. ***Forecast combination*:** Bates and Granger (1969),…

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

8) Empirical Performance of High-Dimensional Methods

(a) Data selection and preparation issues
(b) Comparisons among factor estimation methods
(c) Comparisons among many-predictor forecasting methods
(d) Empirical evidence on in-sample fit of DFM model
(e) Many-predictor methods vs. the world

Disclaimer: There now is a large literature and considerable practitioner experience with empirical DFMs, and a smaller but also substantial literature examining other many-predictor methods. This discussion is informed by this body of empirical knowledge but does not pretend to be a comprehensive survey. See the survey and meta-analysis by Eichmeier and Ziegler (2008) for a bibliography.

(a) Data selection and preparation issues

Bear in mind that…

- The factors you get out depend on the data you put in.

- More variables do not always mean more information, for example putting in CND, CD, CS *and* total consumption doesn't make sense (aggregation identity).

- Judgment should be exercised about the balance between various categories of data; if most of the data are production and output, your dominant factor will be an output factor

(b) Comparisons among factor estimation methods

Discussed above. Empirical evidence suggests estimation method is not a first order issue although there is limited evidence on MLE (2-step or full) to date.

(c) Contemporaneous fit (fraction R^2 explained)

Watson (2004) comment on Giannone, Reichlin and Sala (2004)

Table 1
Fraction of variance explained by one- and two-factor models

	Sargent and Sims[1]		Giannone, Reichlin, and Sala[2]	
Series	1 factor	2 factors	1 factor	2 factors
Average weekly hours	0.77	0.80	0.49	0.61
Layoffs	0.83	0.85	0.72	0.82
Employment	0.86	0.88	0.85	0.91
Unemployment	0.77	0.85	0.74	0.82
Industrial production	0.94	0.94	0.88	0.93
Retail sales	0.46	0.69	0.33	0.47
New orders durables	0.67	0.86	0.65	0.74
Sensitive material prices	0.19	0.74	0.53	0.60
Wholesale prices	0.20	0.69	0.34	0.67
M1	0.16	0.20	0.15	0.30

1. From Table 21 of Sargent and Sims (1977).
2. From Appendix 6.2.

(c) Contemporaneous fit (fraction R^2 explained) (ctd)

Stock and Watson (2005)

- Test exact DFM restrictions, find large fraction of rejections in U.S. quarterly data
- But the rejections are all very small in a R^2 sense.
- The approximate DFM seems to be a good description of the data

(d) Comparisons among many-predictor forecasting methods

Papers include Inoue and Kilian (2008), Koop and Potter (2004), Bańbura, Gianonne, and Reichlin (2008), Stock and Watson (2006a, 2006b, 2009), Pesaran, Pick, and Timmerman (2009).

- DFMs often outperform the many-predictor statistical methods.
- Stock and Watson (2009) conclusions:
  - DFMs provide better forecasts than other many-predictor methods for real series, some interest rates.
  - Other methods e.g. BMA can provide better forecasts by using many principle components for some series – real wages, regional housing construction, etc (no clear unifying pattern)
  - Some series are just hard to forecast and many predictors don't seem to help! Price inflation, stock returns, exchange rates.

(e) Many-predictor methods vs. all other forecasting methods

Factor forecast performance depends on the application

Real variables

US, EU – generally find substantial improvements (especially US) over other models. Improvement over competitors is greater as horizon increases. Improvement is possibly enhanced using some cointegration (observed EC terms, in addition to factors).

Inflation

U.S. survey by Stock and Watson (2008). Performance depends strongly on the sample. Pre-85, factor models worked extremely well. 85-06, very hard to beat random walk forecast (episodic). Current recession is a real-time test of the factor forecasts.

Selected references:

Eichmeier and Ziegler (2008) (metastudy), Stock and Watson (2009)

Summary

1. The quest for exploiting large data sets has made considerable advances
2. Large *n is* a blessing – turning the principle of parsimony on its head ($N^2/T \to \infty$ results)
3. State of knowledge of DFM estimation and factor extraction is pretty advanced: it doesn't seem to make a lot of difference what method you use if *n* is large, but this said the MLE (two-step seems to be enough) has some nice properties theoretically and in initial applications.
4. Applications to forecasting are well advanced and implemented in real time. Applications to SVARs (FAVAR), IV estimation, and DSGE estimation are promising.