High Dimensional Forecasting

Francis J. DiTraglia

University of Pennsylvania

Econ 722

What Have We Learned So Far?

1. Classical Model Selection
2. Focused Model Selection
3. Moment Selection for GMM
4. Shrinkage Estimation
5. Factor Models

Today's Lecture – A Bit of a Grab Bag

"Tie everything together" by looking at high-dimensional forecasting problems, consider some avenues for future research. Also give a brief overview of some topics I had hoped to cover in more detail but didn't have time for.

Some Special Problems in High-dimensional Problems

Estimation Uncertainty

We've already seen that OLS can perform very badly if the number of regressors is large relative to sample size.

Best Subsets Infeasible

With more than 30 or so regressors, we can't check all subsets of predictors making classical model selection problematic.

Noise Accumulation

Large N is supposed to help in factor models: averaging over the cross-section gives a consistent estimator of factor space. This can fail in practice, however, since it relies on the assumption that the factors are *pervasive*. See Boivin & Ng (2006).

Main References

Stock & Watson (2006) – “Forecasting with Many Predictors”

Overview of high-dimesional forecasting with a review of forecast combination, factor models, and Bayesian approaches.

Ng (2013) – “Variable Selection in Predictive Regressions”

Reviews and relates a number of shrinkage & selection methods.

Stock & Watson (2012)

Examines a wide range of shrinkage procedures to see if they can improve on diffusion index forecasts.

Kim & Nelson (2013)

“Horse Race” of various factor and shrinkage methods for forecasting.

Diffusion Index Forecasting – Stock & Watson (2002a,b)

JASA paper has the theory, JBES paper has macro forecasting example.

Basic Setup

Forecast scalar time series y_{t+1} using N-dimensional collection of time series X_t where we observe periods $t = 1, \ldots, T$.

Assumption

Static representation of Dynamic Factor Model:

$$\begin{aligned} y_t &= \beta' F_t + \gamma(L) y_t + \epsilon_{t+1} \\ X_t &= \Lambda F_t + e_t \end{aligned}$$

"Direct" Multistep Ahead Forecasts

"Iterated" forecast would be linear in F_t, y_t and lags:

$$y^h_{t+h} = \alpha_h + \beta_h(L) + \gamma_h(L) + \gamma_h(L) y_t + \epsilon^h_{t+h}$$

This is really just PCR

Diffusion Index Forecasting – Stock & Watson (2002a,b)

Estimation Procedure

1. Data Pre-processing
   1.1 Transform all series to stationarity (logs or first difference)
   1.2 Center and standardize all series
   1.3 Remove outliers (ten times IQR from median)
   1.4 Optionally augment X_t with lags
2. Estimate the Factors
   - No missing observations: PCA on X_t to estimate $\widehat{F}_t$
   - Missing observations/Mixed-frequency: EM-algorithm
3. Fit the Forecasting Regression
   - Regress y_t on a constant and lags of $\widehat{F}_t$ and y_t to estimate the parameters of the "Direct" multistep forecasting regression.

Diffusion Index Forecasting – Stock & Watson (2002b)

Recall from last time that, under certain assumptions, PCA consistently estimates the space spanned by the factors. Broadly similar assumptions are at work here.

Main Theoretical Result

Moment restrictions on (ϵ, e, F) plus a "rank condition" on Λ imply that the MSE of the procedure on the previous slide converges to that of the infeasible optimal procedure, provided that $N, T \rightarrow \infty$.

Diffusion Index Forecasting – Stock & Watson (2002a)

Forecasting Experiment

- Simulated real-time forecasting of eight monthly macro variables from 1959:1 to 1998:12
- Forecasting Horizons: 6, 12, and 24 months
- "Training Period" 1959:1 through 1970:1
- Predict h-steps ahead out-of-sample, roll and re-estimate.
- BIC to select lags and # of Factors in forecasting regression
- Compare Diffusion Index Forecasts to Benchmark
  - AR only
  - Factors only
  - AR + Factors

Diffusion Index Forecasting – Stock & Watson (2002a)

Empirical Results

- Factors provide a substantial improvement over benchmark forecasts in terms of MSPE
- Six factors explain 39% of the variance in the 215 series; twelve explain 53%
- Using all 215 series tends to work better than restricting to balanced panel of 149 (PCA estimation)
- Augmenting X_t with lags isn't helpful

What about Ridge and Lasso?

Basic Idea

Diffusion index forecasts are really just PCR. Why not try Ridge or Lasso with all predictors rather than estimating factors?

De Mol, Giannone & Reichlin (2008)

- Compare PCA-based factor forecasts to Ridge and Lasso
- In a small out-of-sample experiment, Ridge and Lasso with appropriate penalty parameters give results comparable to diffusion index.
- Analyze asymptotics of Ridge under assumptions typically used to justify PCA

Other Ways of Extracting Factors

Sparse PCA

Independent Components Analysis (ICA)

Targeted Factors?

Notice that PCA (and indeed ICA and sparse PCA) completely ignore Y when extracting factors!

Bai and Ng targeted predictors paper adds transformations of original series. (Talk more about this in relation to kernel stuff). To target or not to target? Why should the PCs of X be related to Y? Lots of ways to target. Bai and Ng as well as PLS Groen & Kapetanios, Kelly & Pruitt. Sparse PCA, ICA.

To Target or Not to Target?

Problem with PCA and Friends

Completely ignores Y in constructing the factors!

$$X_t \quad =$$

Partial Least Squares (PLS)

As an Optimization Problem

Construct a sequence of linear combinations of X that solve

$$\max_{\alpha} \mathit{Corr}^2(\mathbf{y}, X\alpha)\mathit{Var}(X\alpha)$$

subject to $||\alpha|| = 1$ and the constraint that each PLS “factor” is orthogonal to the preceding ones.

As a Probabilistic Model

Bootstrap Aggregation – "Bagging"

Bagging Algorithm

1. Make a bootstrap draw
2. Carry out selection/shrinkage/estimation using boostrap data
3. Use estimated parameters from to construct a forecast $\widehat{y}^{(b)}_{T+h}$
4. Repeat for $b = 1, \ldots, B$
5. Average to get "Bagged" Forecast: $\widehat{y}^{(Bag)}_{T+h} = \frac{1}{B}\sum_{b=1}^{B}\widehat{y}^{(b)}_{T+h}$

Details

- If the data are dependent, need block bootstrap.
- In step 3, we forecast using the *parameters* estimated from the bootstrap data but the *predictors* from the *real* dataset.

Bootstrap Aggregation – “Bagging”

Why Bagging?

- Aims to reduce the forecast error of “unstable” procedures such as variable selection of Lasso, by reducing their variance.
- Completely portable: you can bag *anything* provided you have an appropriate way to carry out the bootstrap.
- May provide a way of attacking the problem of inference post-model selection. See Efron (JASA, *Forthcoming*) “Estimation and Accuracy after Model Selection”

Bagging in Economics

Inoue & Killian (2008, JASA)

Compares performance of bagged "pre-test" estimator (variable selection via a t-test) to other methods of forecasting US Inflation. Bagging is carried out via a block bootstrap.

Stock & Watson (2012)

Among other shrinkage procedures, they consider a large-sample approximation to bagging pre-test estimators that doesn't require making bootstrap draws.

Other Papers That Use Bagging

- Hillebrand & Medeiros (2010): Realized Volatility Forecasts
- Hillebrand et al (2012): Forecasting the Equity Premium
- Kim and Swanson (2013)

Boosting

Ensemble Methods

Machine learning term for "non-Bayesian model averaging"

What is Boosting?

- Combine large number of "weak learners" (i.e. crappy predictive models) so that the *ensemble* predicts well.
- Explicitly designed around predictive loss
- Arbitrarily improve in-sample fit of arbitrarily the weak learners!

Book-Length Treatment

Shapire & Freund (2012) – *Boosting: Foundations and Algorithms*

Bai & Ng (2009) – Boosting Diffusion Indices

Also, Buchen & Wohlrabe (2011) – boosting compared to the Stock and Watson 2006 example.

We'll learn more about boosting in the student presentations!

Some Avenues to Explore

- Kernel Methods (Exterkate et al, 2013)
- Model selection for factor / shrinkage estimators
  - Generated regressor problem
  - Choice of Lasso and Ridge Penalties in time-series Forecasting
- When and how should we extract targeted factors?
- Can we provide an economic or statistical justification for sparse PCA or ICA?