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.

What's Different About High-Dimensional Problems?

- OLS performs very badly if the number of regressors is large relative to sample size.
- Estimation uncertainty can be a problem
- Noise accumulation in PCA: that Fan and Li paper, also Boivin & Ng

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)

Kim & Nelson (2013)

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:

$$y_{t+h}^h = \alpha_h + \beta_h(L) + \gamma_h(L) + \gamma_h(L) y_t + \epsilon_{t+h}^h$$

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 $N, T \to \infty$.

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

Diffusion Index is really just PCR. What about Ridge and Lasso?
Rather than extract factors, just do penalized forecasting regression and "toss everything in."
De Mol, Giannone & Reichlin (2008) – Compare performance of Ridge and Lasso to PCA-based forecasts (PCR). Small out-of-sample experiment Same data as Stock & Watson 2005 – implications of Dynamic Factors for VAR analysis). With appropriate choice of penalty parameters, Ridge and Lasso forecasts performance similar to that of Diffusion Index (PCR).
Analyze asymptotic behavior of ridge under the assumptions used to justify PCA by Stock & Watson inter alia.
Talk more about this looking at the two more recent papers.

Other Ways of 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.

Kernel Methods

Hasn't been explored much in econometrics, but very popular in machine learning. One recent econometrics paper is Exterkate et al (2013) "Nonlinear Forecasting with Many Predictors Using Kernel Ridge Regression"

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 $\widehat{y}_{T+h}^{(b)}$
4. Repeat for $b = 1, \ldots, B$
5. Average to get "Bagged" Forecast: $\widehat{y}_{T+h}^{(Bag)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{y}_{T+h}^{(b)}$

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!

What Should We Make of This Literature

- Shrinkage is essential in high-dimensional problems and often beats model averaging.
- There's no single best procedure for every problem
- Worth considering new ways of extracting factors & combining with shrinkage
- How and when should we target? Relationship to spurious correlation and overfitting? There are strong opinions on both sides, but I haven't seen any compelling justifications.
- No one has any idea what's going on with model selection for these problems: generated regressors, target-specific, consistency vs. efficiency. Onatski paper is one contribution but it deals with a somewhat odd version of the problem: forecasting the whole X_t
- Want shrinkage to be data-driven. There are results on rates for consistency of Lasso, etc, but that's not what we care about for forecasting. People just use the default CV procedure in matlab (or R) which has no justification in the time-series setting.
- Two-step procedures, generated regressors.