Forecasting Economic Variables with High-dimensional Time Series

Christine De Mol

Université Libre de Bruxelles
Dept Math. et ECARES

Workshop on
"Inverse problems: theory and statistical inference"
International Academic Forum Heidelberg
October 28-29, 2016

Forecasting

"Prediction is very difficult, especially about the future"
(attributed to Niels Bohr)

Forecasting by combining/aggregating information

Joint work with Domenico Giannone (FED New York) and Lucrezia Reichlin (London Business School)

- Let X_t be a high-dimensional time series or "panel" of time series, with (cross-sectional) dimension N, e.g. of N macroeconomic or financial variables, observed at discrete time intervals $t = 1, 2, \ldots,$ e.g. every day, month, quarter or year.
- Assume that each individual time series in the panel is a stationary process (or has been transformed to be so), having mean zero and unit variance.
- Goal: forecast a given economic variable y_t (included or not in the panel), e.g. inflation, unemployment or GDP growth, based on the information contained in the whole panel, and not only based on the past of y_t ("Big Data" instance).

Forecasting high-dimensional time series

- "Input" (data) matrix: $\{x_{nt}\}$, for $n = 1, 2, \ldots, N$ and $t = 1, 2, \ldots, T$, to be ranged in a T by N matrix X.
- Wanted: y_{T+h}, variable to be forecast at a given time horizon h, on the basis of X, i.e. of the information available at time T.
- "Output" (response): y_{t+h} for each $t = 1, 2, \ldots, T-h$ ("supervised" setting).
- Assume linear dependence: $y_{t+h} = \sum_n \beta_n x_{nt}$ or $y = X\beta$, where $y = (y_{1+h}, y_{2+h}, \ldots, y_T)'$ (prime = transpose)
- To simplify, without loss of generality, we can write all formulas for $h = 0$ (equivalent to a proper redefinition of the time labels)
- Include, if necessary, the p lagged time series $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ into the panel, as in VAR (Vector AutoRegressive) models (and redefine N accordingly).

Two distinct problems

- Linear regression problem: $y = X\beta$.
- Prediction ("generalization"):
  predict (forecast) the response y.
- Identification (Variable Selection):
  find the coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_N)'$
  i.e. identify the most relevant predictors

  or SELECT them when many coefficients are zero
  i.e. when β is SPARSE.

  Essential for interpretation!

Ordinary Least-Squares (OLS) Regression

- Noisy data: $y = X\beta + e$ (e = zero-mean error term)
- Reformulate problem as a classical linear regression problem: minimize quadratic loss function

$$\Lambda(\beta) = \|y - X\beta\|_2^2 \qquad \left(\|y\|_2 = \sqrt{\sum_t |y_t|^2} = L_2\text{-norm}\right)$$

- Equivalently, solve variational (Euler) equation

$$X'X\beta = X'y$$

- If $X'X$ is full-rank, minimizer is OLS solution

$$\beta_{ols} = (X'X)^{-1}X'y$$

Problems with OLS

- Not feasible if $X'X$ is not full-rank i.e. has eigenvalue zero (in particular, whenever $N > T$). In many practical cases, $N >> T$ (as for short macroeconomic time series) (large p, small n paradigm).
- Then the minimizer is not unique (system largely underdetermined), but you can restore uniqueness by selecting the "minimum-norm least-squares solution", orthogonal to the null-space of X (OK for prediction but not necessarily for identification!).
- Also $X'X$ may have eigenvalues close to zero (happens when both N and T get large) $\longrightarrow$ $X'X$ has a large "condition number" (= ratio between largest and smallest e.v.). This is ill-conditioning, also referred to as "curse of dimensionality".

Principal Component Regression

- For high-dimensional time series, the standard remedy is Principal Component Regression (PCR) (Stock and Watson 2002 , for static PC; Forni, Hallin, Lippi, Reichlin 2000, for dynamic PC, i.e. PC in Fourier):

$$\beta_{pcr} = \sum_{k=1}^{K} \frac{\langle X'y, v_k \rangle}{d_k^2} \; v_k$$

  where v_k are the eigenvectors of $X'X$ with eigenvalues d_k^2 (="Truncated SVD", at some K, smaller than the true rank).
- This is a kind of "regularization"($<$ inverse problem theory) providing the necessary dimension reduction to avoid ill-conditioning ($\rightarrow$ extreme volatility of estimators) by introducing bias to decrease variance.
- Alternative to the standard PCR paradigm in economics: penalized regression (ridge, lasso, etc.) (De Mol, Giannone, Reichlin 2008, and also Banbura, Giannone, Reichlin 2010, for "Bayesian VARs").

Alternative: penalized regression

- Minimize the least-squares residual augmented by a penalty (with a tuning parameter called the "regularization parameter").
- Ridge regression (Hoerl and Kennard 1970 ; Tikhonov's regularization in inverse problem theory):
  penalize the (squared) L_2-norm of β: $\|\beta\|_2^2 = \sum_{n=1}^{N} |\beta_n|^2$
  NB. Quadratic penalties provide solutions (estimators) which depend linearly on the response y but do not allow for variable selection (typically all coefficients are different from zero).
- Lasso regression (Tibshirani 1996):
  penalize L_1-norm of β: $\|\beta\|_1 = \sum_{n=1}^{N} |\beta_n|$
  NB. Enforces sparsity of β, i.e. the presence in this vector of many zero coefficients $\longrightarrow$ Variable selection is performed!

Bayesian framework

- OLS can be viewed as maximum (log-)likelihood estimator for gaussian "noise"
  $\rightarrow$ penalized maximum likelihood.
- Bayesian interpretation: MAP estimator and penalty interpreted as a prior distribution for the regression coefficients.
- Ridge $\sim$ Gaussian prior.
- Lasso $\sim$ Laplacian prior (double exponential).

Turning the curse of dimensionality into blessing

- What can we learn from the data?
- (Macroeconomic) series are highly correlated; lots of comovement.
- Does the accumulation of data and of series help forecasting the target variable?
- Asymptotics for $T \to \infty$ and $N \to \infty$?

Ridge regression

- Linear regression model

$$y = X\beta + e$$

- Ridge estimator

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{T} \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

i.e.

$$\hat{\beta} = \left(\frac{X'X}{T} + \lambda I \right)^{-1} \frac{X'y}{T}$$

- Under fairly general (standard) assumptions, we have that $\frac{\|X'e\|}{\sqrt{NT}} = O_p(1)$ as $N, T \to \infty$

Ridge regression

- Root Mean Square Forecast Error (RMSFE)

$$\frac{1}{\sqrt{T}}\|X\hat{\beta} - X\beta\| \leq \sqrt{\lambda}\|\beta\| + \frac{1}{\sqrt{\lambda}} O_p(\sqrt{\frac{N}{T}})$$

- Optimal value of λ (minimizing the bound, i.e. equally balancing the two terms)

$$\lambda \sim \frac{\sqrt{N}}{\sqrt{T}\|\beta\|}$$

- $\longrightarrow$ asymptotic rate for the RMSFE

$$\frac{1}{\sqrt{T}}\|X\hat{\beta} - X\beta\| \sim \frac{N^{1/4}}{T^{1/4}}\sqrt{\|\beta\|}$$

Factor models for high-dimensional time series

- A way of modelling strong comovement, assuming that the panel is driven by a small number of factors spanning a subspace of (fixed) dimension K:

$$X_t = \Lambda F_t + \xi_t \qquad \text{(common + idiosyncratic)}$$

- Assumptions:
  1. the factors F_t are a K-dimensional stationary process, with covariance matrix $\mathrm{E} F_t F_t' = I_K$;
  2. the residuals ξ_t are a N-dimensional stationary process, orthogonal to the factors, with covariance matrix $\mathrm{E}\xi_t\xi_t' = \Psi$ (of full-rank, for every N);
  3. the matrix Λ loading the factors is a non-random matrix of dimension $N \times K$ and of full-rank K for every N;
  4. all eigenvalues of $\Lambda'\Lambda$ grow as N (all predictors are informative on the factors).
- Then the population covariance matrix is $\Sigma_{XX} = \mathrm{E}(X_t X_t') = \Lambda\Lambda' + \Psi$ (here $\Psi = I_N$, for simplicity) and has two clusters of eigenvalues with a spectral gap.

Consistency and rates for $T \to \infty$ and $N \to \infty$

(De Mol, Giannone, Reichlin 2008, and work in progress)

- Under the assumptions above (factor model), one can show that $\|\beta\| \sim \frac{1}{\sqrt{N}}$, yielding a decay rate $T^{-1/4}$ for the RMSFE.

- Classical asymptotics: N fixed, $T \to \infty$.
  Then, assuming OLS feasible (i.e. the smallest eigenvalue of $\frac{X'X}{T}$ bounded from below by a positive constant c), and setting $\lambda \sim \frac{1}{\sqrt{T}}$ (or even $\lambda = 0$, i.e. without regularization), one gets a rate T^{-1} for the RMSFE (and $T^{-1/2}$ for the estimation of β).

Consistency and rates for $T \to \infty$ and $N \to \infty$

- Taking into account the spectral gap and using Weyl's perturbation Lemma to control the difference between the eigenvalues of the population and sample covariance matrices, and with $\lambda \sim \frac{N}{\sqrt{T}}$, we can show that

$$|X_t'\hat{\beta} - X_t'\beta| \leq O_p(\frac{1}{\sqrt{N}}) + O_p(\frac{1}{\sqrt{T}})$$

  which shows consistency along any path in (N, T).
- Same rates as for PCR, when knowing K
  NB. Estimating the "true" number of factors K is a hard problem, widely discussed in the literature.

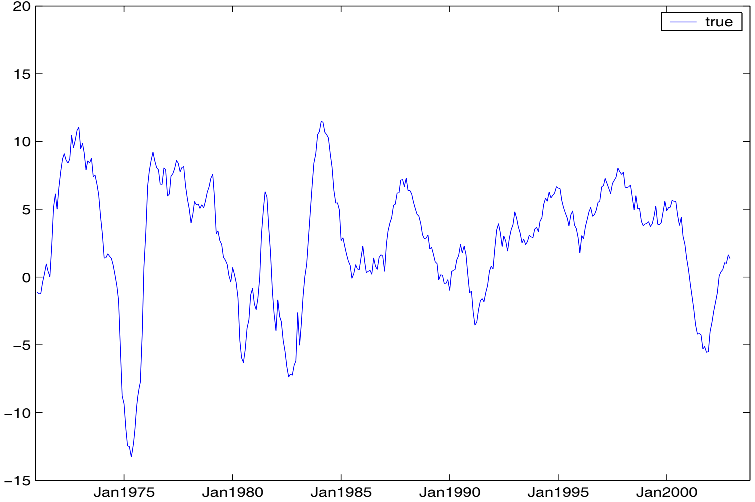
Generalizations

- Extension to the case where the idiosyncratic components are mildly correlated ($\Psi \neq I$) ("approximate" factor models), so that the largest eigenvalue of Σ_{XX} in the second cluster of Σ_{XX} can grow with N but at a slower rate than N. Consistency still holds but at a slower rate.
- Extension to non-uniform and non i.i.d. priors on the components of β: $\beta \sim \mathcal{N}(0, \Phi)$.
- Extension to other linear regularization method based on "spectral filtering".

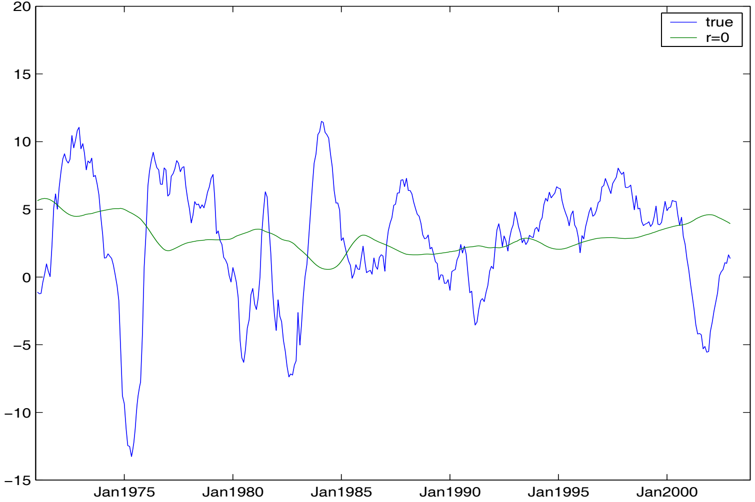
Macroeconomic forecasting: empirical results

- Macroeconomic data-set of 131 monthly time series for the US economy from Jan59 to Dec03 (Stock and Watson 2005), transformed for stationarity and standardized.
- Variable to forecast:
  1. Industrial Production: $y_{t+h} = (\log IP_{t+h} - \log IP_t) \times 100$
  2. Price inflation: $y_{t+h} = \pi_{t+h} - \pi_t$
- Simulated out-of-sample exercise:
  For each time $T = \mathrm{Jan}70, ..., \mathrm{Dec}01$, estimate β using the most recent 10 years of data (rolling scheme),
  with a forecast horizon of $h = 12$ months.
  (No lags of the regressors included here;
  similar results when including lags)

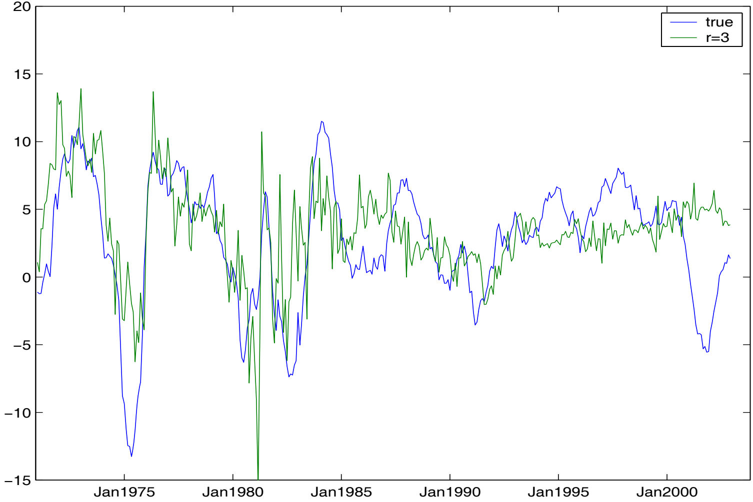
Forecasting IP (actual series)



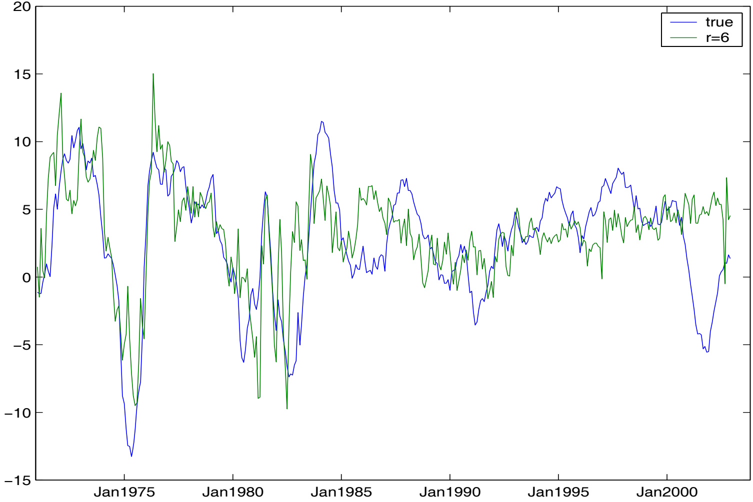
Forecasting IP (PCR; $K = 0$, naive RW forecast)



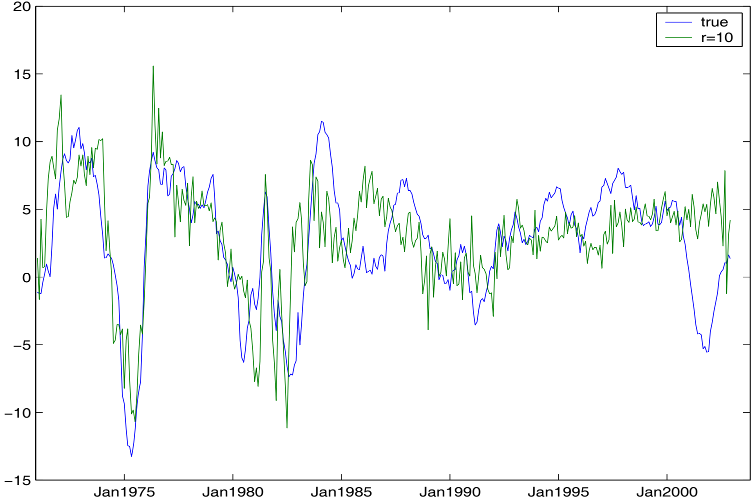
Forecasting IP (PCR; $K = 3$)



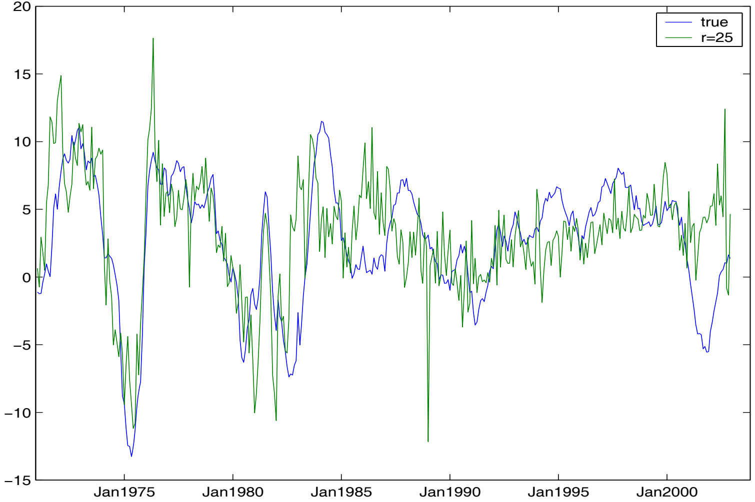
Forecasting IP (PCR; $K = 6$)



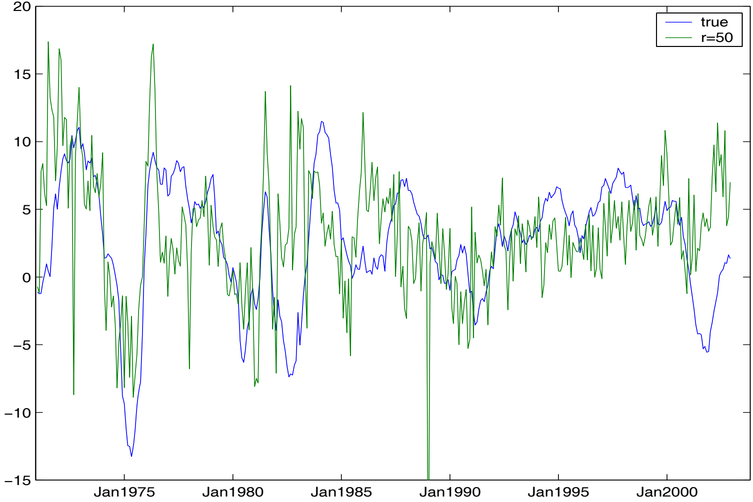
Forecasting IP (PCR; $K = 10$)



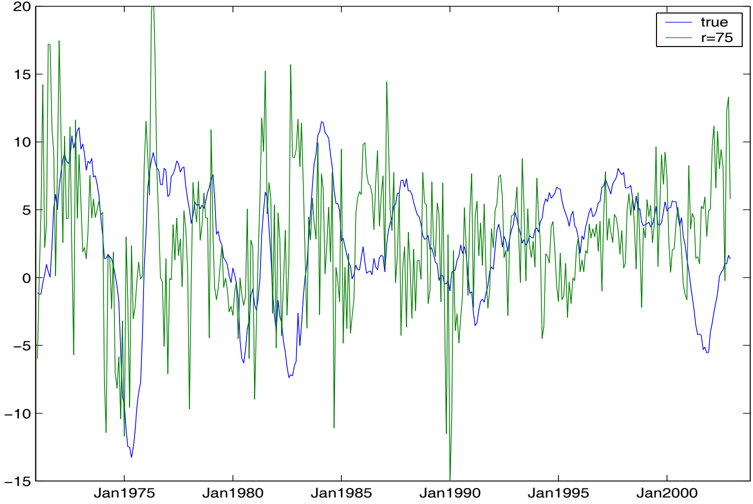
Forecasting IP (PCR; $K = 25$)



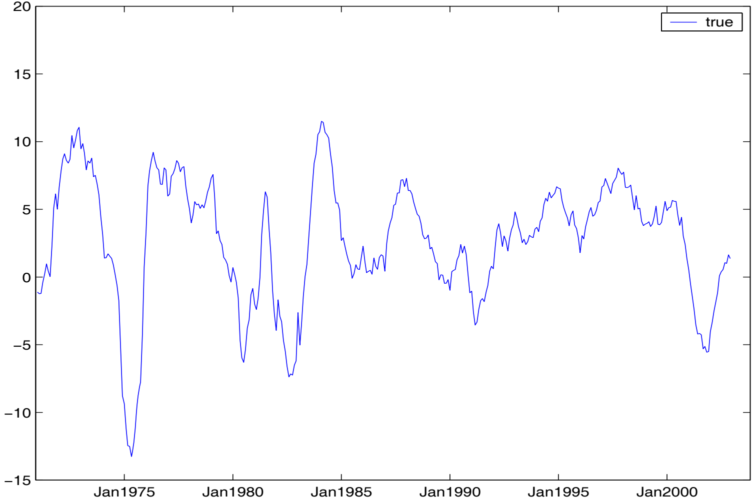
Forecasting IP (PCR; $K = 50$)



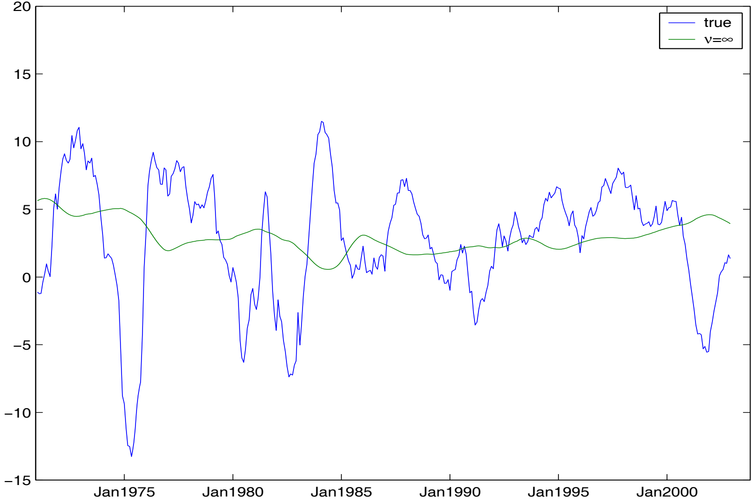
Forecasting IP (PCR; $K = 75$)



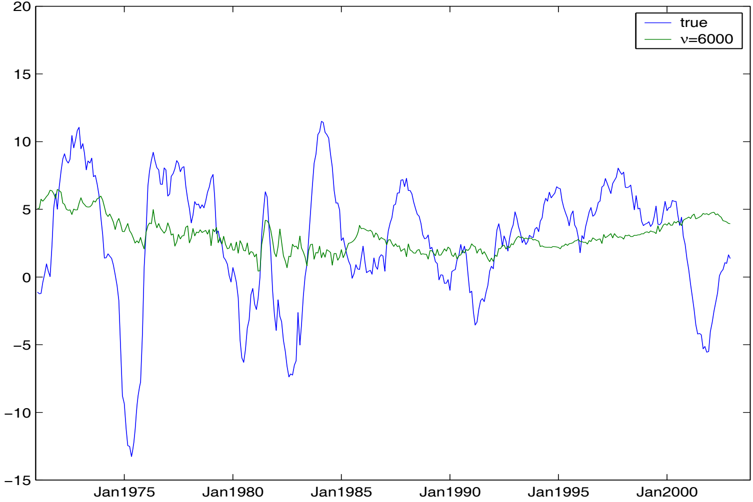
Forecasting IP



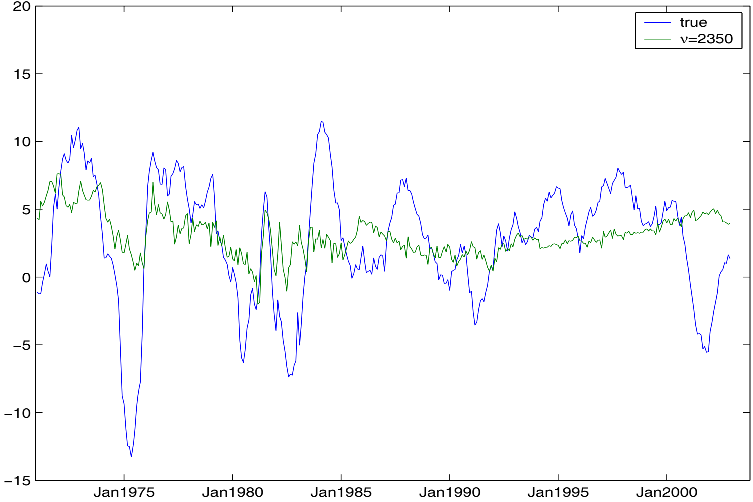
Forecasting IP (ridge)



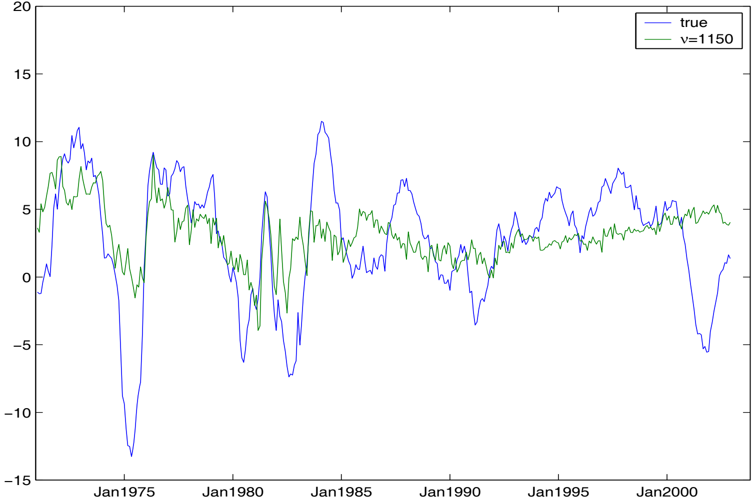
Forecasting IP (ridge)



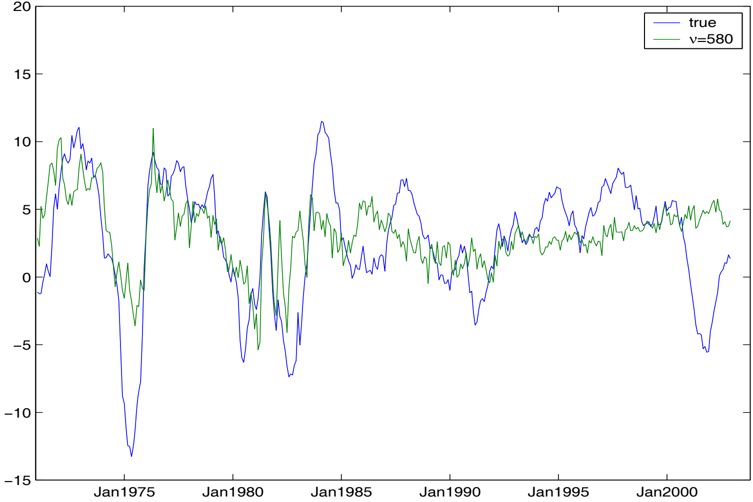
Forecasting IP (ridge)



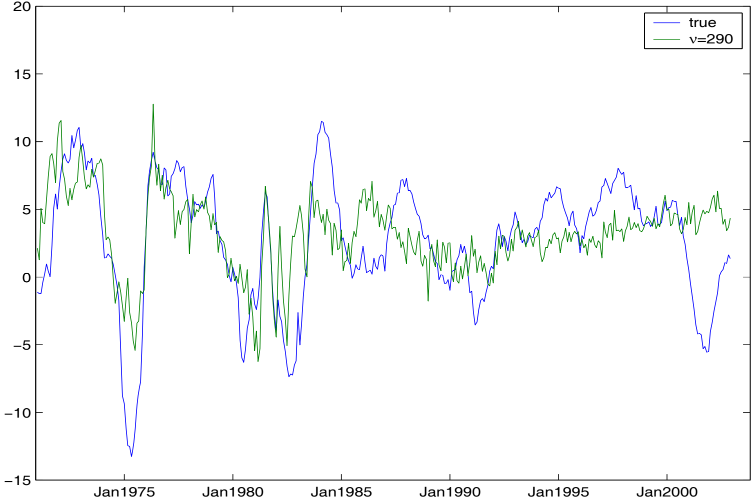
Forecasting IP (ridge)



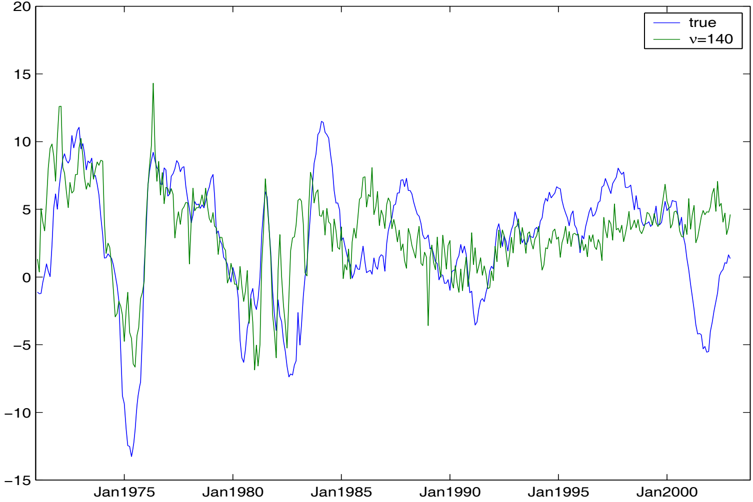
Forecasting IP (ridge)



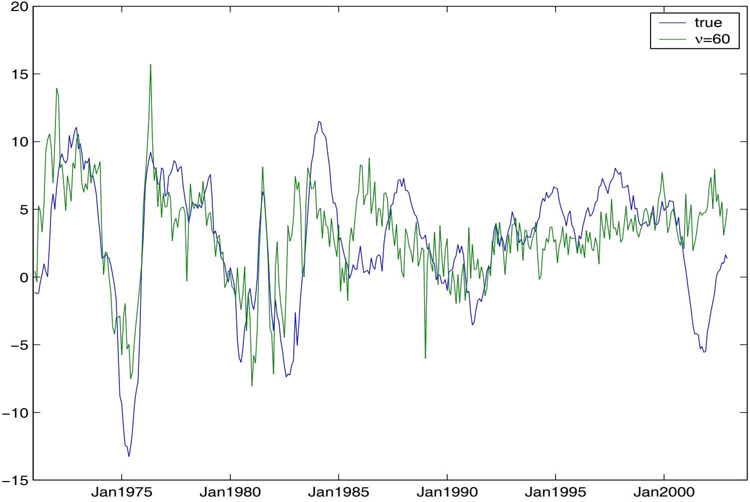
Forecasting IP (ridge)



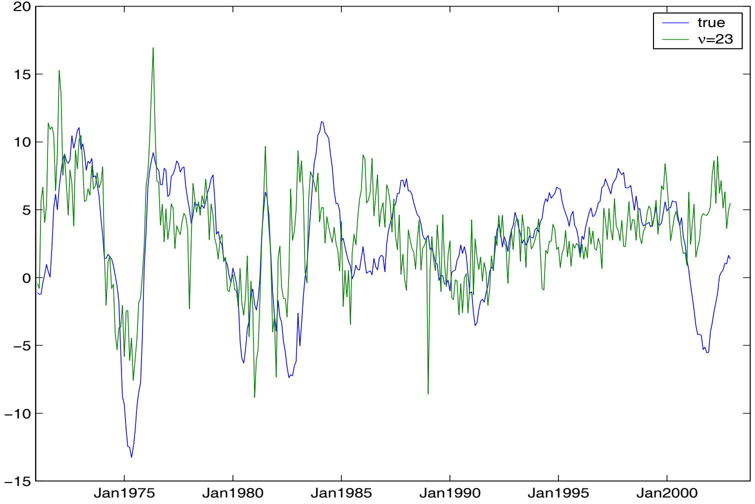
Forecasting IP (ridge)



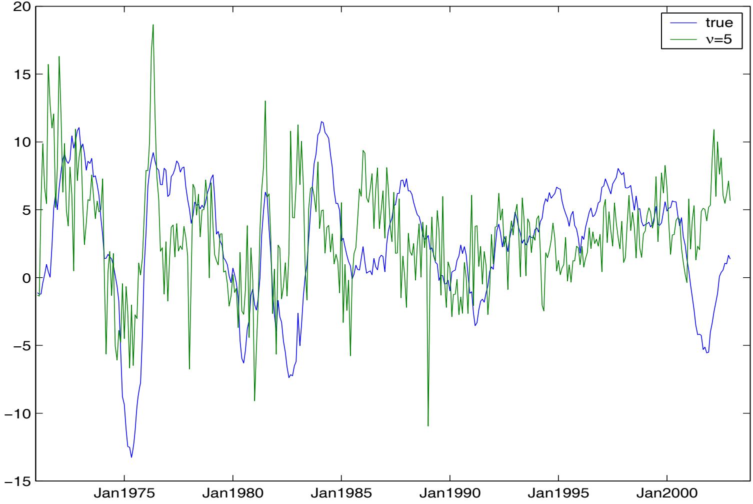
Forecasting IP (ridge)



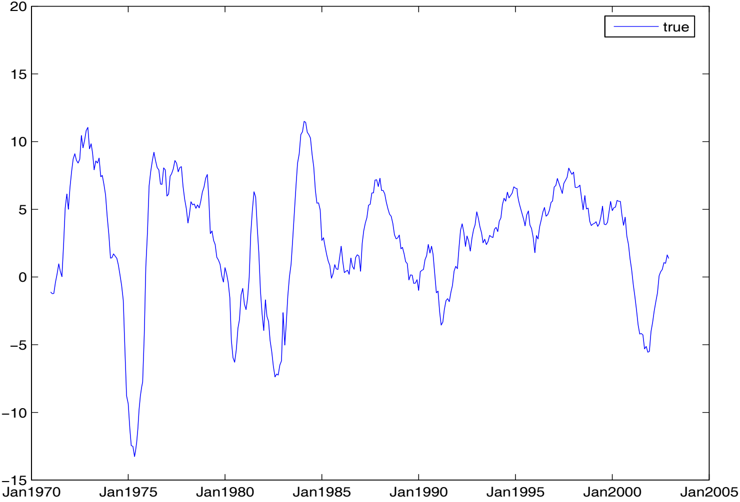
Forecasting IP (ridge)



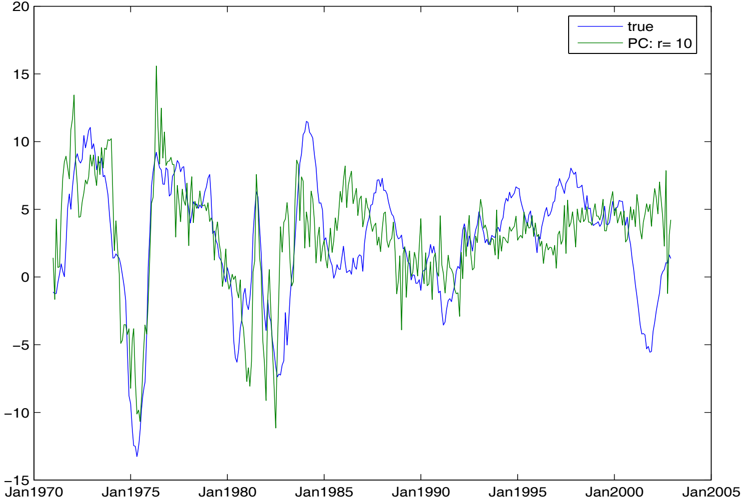
Forecasting IP (ridge)



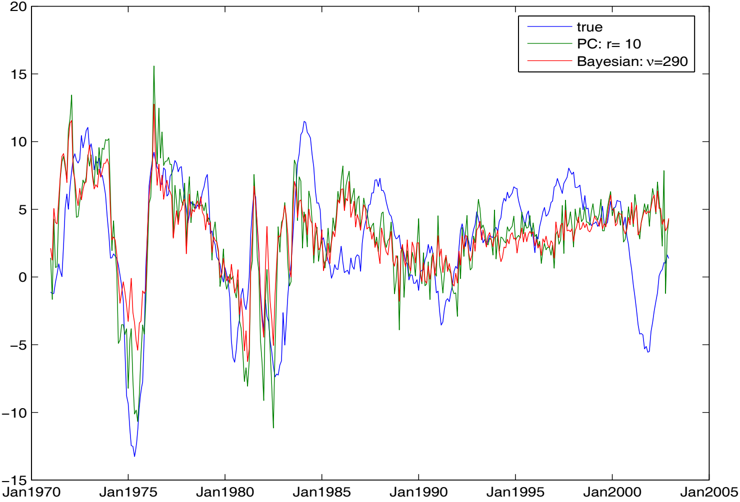
Forecasting IP



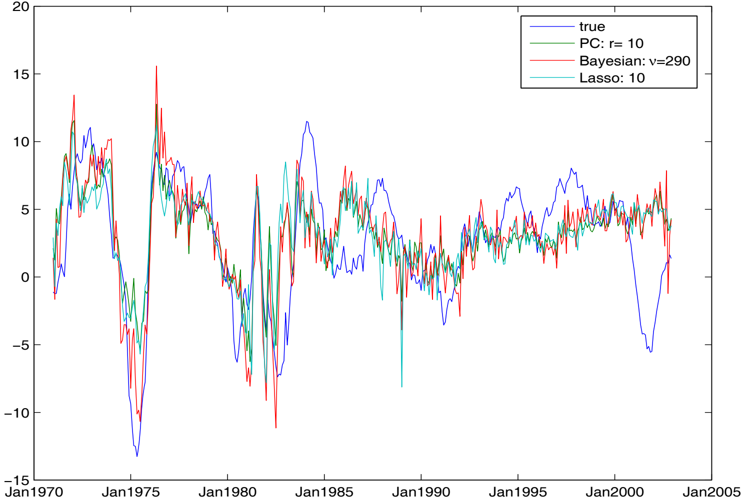
Forecasting IP (PCR; least MSFE, $K = 10$)



Forecasting IP (ridge vs PCR; least MSFE)



Forecasting IP (ridge and lasso vs PCR; least MSFE)



Remarks

- In the previous exercice, the number of PC, the ridge and the lasso regularization parameters have been assessed by cross-validation, in order to minimize, when using the rolling scheme, the (out-of-sample) MSFE over the available historical period.
- The ridge parameter can also be set on the basis of the asymptotic consistency results.
- Factor models are a paradigm for high-dimensional time series with a lot of comovement.
  NB. *"Essentially, all models are wrong, but some are useful" (George Box)*
- We can (asymptotically) learn the subspace spanned by the factors and capture the bulk of variation in the panel for forecasting purposes.

Forecast combination

Joint work (2015) with Cristina Conflitti (Bank of Italy) and Domenico Giannone (FED New York)

- To improve accuracy, combine (linearly by means of time-independent weights) individual forecasts $\hat{y}_{i,t+h}$ of the variable y_{t+h} provided by different sources (professional forecasters, different series, models, etc.) $\sum_{i=1}^{N} w_i \hat{y}_{i,t+h} \equiv \mathbf{w}' \hat{\mathbf{y}}_{t+h}$
- Optimal weights: minimize the mean square forecast error (assuming that the variable y_t is observed for $t = 1, \ldots, T$)

$$\hat{\mathbf{w}}_{\text{OPT}} = \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{t=1}^{T-h} (y_{t+h} - \mathbf{w}' \hat{\mathbf{y}}_{t+h})^2 \text{ s.t. } \sum_{i=1}^{N} w_i = 1 \text{ and } w_i \geq 0$$

- Take $(\mathbf{w}_{\text{OPT}})' \hat{\mathbf{y}}_{T+h}$ to forecast y_t at time $t = T + h$

Forecast Combination as Portfolio Optimization

- Similar problem for Markowitz portfolios: find minimum-variance (i.e. without target-return constraint) no-short (i.e. with nonnegative weights) portfolios, the vector of forecasts being replaced with a vector of returns.
- Special case of a “lasso” regression (Tibshirani, 1996) since the two constraints, meaning that the weight vector belongs to the unit (probability) simplex, imply that the weight vector has a unit L_1-norm: $\sum_{i=1}^{N} |w_i| = 1 \longrightarrow$ the weight vector is sparse (contains many zeroes).
- But sparsity is not “tunable” : “parameter-free regularization” (drawback?)
- Special case of the “Sparse and stable Markowitz portfolios” (with tunable sparsity), for which a constrained LARS algorithm was also developed (Brodie, Daubechies, De Mol, Giannone, Loris 2009)

Optimal Combination of Density Forecasts

- Combine individual probability density forecasts

$$\hat{p}(\cdot) = \sum_{i=1}^{N} w_i \hat{p}_i(\cdot) \equiv \mathbf{w}'\hat{\mathbf{p}}(\cdot)$$

  with weight vector in the unit simplex.
- Optimal weights: maximize the "log predictive score"

$$\hat{\mathbf{w}}_{\mathrm{OPT}} = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{1}{T-h} \sum_{t=1}^{T-h} \ln \hat{p}(y_{t+h})$$

- Why? Minimize Kullback-Leibler Information Criterion measuring similarity between between true $p(\cdot)$ and combined density $\hat{p}(\cdot)$

$$\text{KLIC} = \int p(y) \ln \frac{p(y)}{\hat{p}(y)} dy = E[\ln p(\cdot) - \ln \hat{p}(\cdot)]$$

  or its sample average. When reference target density $p(\cdot)$ is missing (as for survey data), notice that the first term yields a constant independent of the weights.

An iterative algorithm to maximize the log score

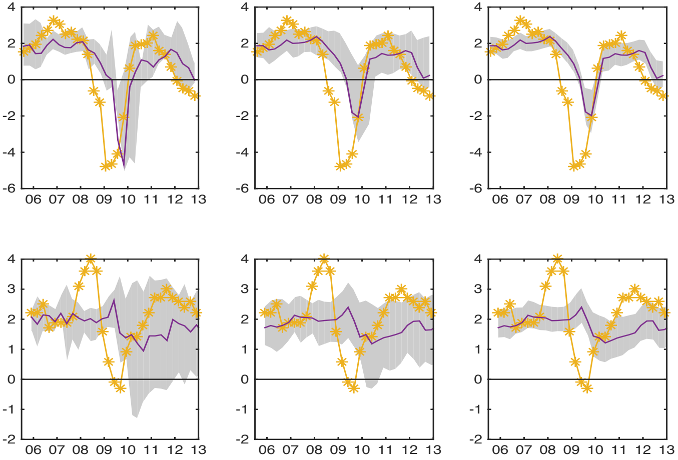
- The following simple multiplicative algorithm

$$w_i^{(k+1)} = w_i^{(k)} \frac{1}{T-h} \sum_{t=1}^{T-h} \frac{\hat{p}_i(y_{t+h})}{\sum_{j=1}^{N} \hat{p}_j(y_{t+h}) w_j^{(k)}}$$

  initialized with positive weights summing to one, e.g. with $w_i^{(0)} = 1/N$, preserves nonnegativity at each iteration.
- derived as a MM-algorithm (through surrogates) $\rightarrow$ monotonic increase of the cost function and convergence of the iterates $\mathbf{w}^{(k)}$ to the maximizer of the log score on the unit simplex.

Survey of Professional Forecasters (ECB)

1st row: GDP growth; 2nd row: (HICP) Inflation (orange=outturn)



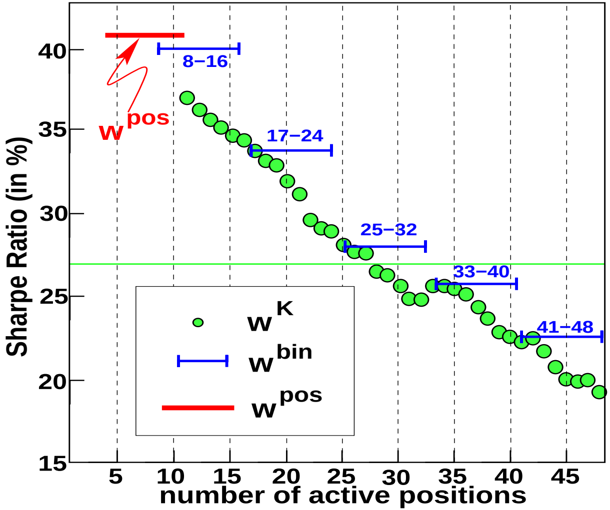
1st column: OPT; 2nd column: EQW; 3rd column: ECB

(purple: combined point forecasts; shaded areas 68% bands)

The $1/N$ Puzzle

- Equal-weight averaging (with weights $1/N$) is very hard to beat!
- Also true for portfolio optimization ("Talmudic portfolios") (DeMiguel, Garlappi and Uppal 2007)
- But here sparsity can help!
  The sparse and stable Markowitz portfolios of Brodie, Daubechies, De Mol, Giannone, Loris (2009) outperform the $1/N$ benchmark in terms of Sharpe ratio $S = m/\sigma$ (m=out-of-sample monthly mean return, σ = standard deviation), at least on two standard benchmark datasets: Fama and French 48 industry portfolios (FF48) and 100 portfolios formed on size and book-to-market (FF100)

Empirical results FF48



Empirical results FF100