Near-equivalence in Forecasting Accuracy of Linear Dimension Reduction Methods in Large Macro-Panels

Efstathia Bura[1]
(joint with Alessandro Barbarino[2])

[1]Applied Statistics, Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology

[2]Federal Reserve Board

October 6, 2017

Outline

Overview of Empirical Results

The Empirical Forecasting Exercise

- McCracken and Ng (2015) created the FRED-MD macro-panel:
  - 123 monthly macro variables.
  - From January 1960 to 2015, balanced panel of 110 predictors with approx. 660 data.
- Focus on: CPI inflation, non-farm payrolls, unemployment rate, labor force participation, average hourly earnings in goods-producing industries, industrial production, real personal consumption expenditures and real personal income.
- Use popular statistical learning methods OLS, PCR, RIDGE, PLS and SIR (Sliced Inverse Regression).
  - Out-of-sample forecasts using **recursive window** at horizons $h = 1, 3, 6, 12$.
  - Compare results in classic horse-race against AR(4).

DFM Forecasting

- Dynamic Factor Models (DFMs) are the mainstream methodology in Econometrics for both measuring comovement and forecasting time series.

- The modeling comprises of two unconnected steps: The factors that replace $\mathbf{x}_t$ are ordered according to how much variance in $\mathbf{X}$ they explain; i.e., in relevance to $\mathbf{X}$ and not to Y.

Best Estimators Before the Great Recession

Table: Recursive Out-of-Sample Window from 1992 to 2007, MSFE Relative to AR4

	h=1		*h=3*		*h=6*		*h=12*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	RIDGE^{b}-119	0.87	RIDGE^{b}-141	0.95	R70SIRb^{c,d}-2	0.95	R8SIRb^{c,d}-2	0.9
PAYEMS	R30SIRb^{c,d}-1	0.96	R30SIRb^{c,d}-1	0.94	R30SIRb^{c,d}-2	0.87	R30SIRb^{c,d}-2	0.82
UNRATE	PC-1	0.93	PLS-1	0.76	RIDGE^{b}-0.4	0.76	RIDGE^{b}-0.3	0.78
CLF16O	RIDGE^{b}-141	0.94	PC-16	0.87	PC-23	0.81	PC-16	0.74
CPIAUC	PC.BS^{h}-8.7	0.89	SIRb^{d}-2	0.94	PC.ON^{f}-1.2	0.97	R8SIRb^{c,d}-2	0.95
CES060	PLS-1	0.99	PLS-1	1	R8SIRb^{c,d}-1	0.97	AR4	1
DPCER	AR4	1	AR4	1	AR4	1	AR4	1
RPI	R8SIRb^{c,d}-1	0.93	R8SIRb^{c,d}-1	0.88	R8SIRb^{c,d}-1	0.88	RFSIRb^{c,d}-1	0.83

Best Estimators In the Recovery

Table: Recursive Out-of-Sample Window from 2010 to 2016, MSFE Relative to AR4

	h=1		*h=3*		*h=6*		*h=12*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	RIDGE^{b}-949	0.96	RIDGE^{b}-949	0.93	RIDGE^{b}-3532	0.94	R30SIRb^{c,d}-3	0.74
PAYEMS	RIDGE^{b}-141	0.74	RIDGE^{b}-141	0.56	RIDGE^{b}-288	0.5	RIDGE^{b}-288	0.44
UNRATE	PCF.BS^{i}-11	0.85	PLS-4	0.64	SIRa^{d}-7	0.55	PC.BS^{h}-1	0.38
CLF16O	PC-7	0.78	PCF.BS^{i}-11	0.52	RIDGE^{b}-0.6	0.39	PLS-5	0.33
CPIAUC	PC.BS^{h}-1	0.96	SIRb^{d}-1.3	0.9	R70SIRb^{c,d}-3	0.94	PC-5	0.87
CES060	SIRa^{d}-7	0.94	SIRb^{d}-1	0.99	R8SIRb^{c,d}-3.5	0.94	R8SIRa^{c,d}-2	0.87
DPCER	PC-5	0.87	PC-5	0.7	PC-5	0.48	PC-17	0.44
RPI	R8SIRa^{c,d}-1	0.91	RFSIRa^{c,d}-1.6	0.79	PLS-8	0.78	PLS-20	0.64

Best Estimators in Full Sample

Table: Recursive Out-of-Sample Window from 1992 to 2016, MSFE Relative to AR4

	h=1		*h=3*		*h=6*		*h=12*	
Target	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*	*Estimator*	*MSFE*
INDPRO	PC-15	0.95	PC-15	0.94	SIRb^{d}-1.4	0.96	R70SIRb^{c,d}-3	0.94
PAYEMS	R8SIRb^{c,d}-1	0.88	R8SIRb^{c,d}-1	0.82	R8SIRb^{c,d}-1	0.89	R30SIRb^{c,d}-2	0.9
UNRATE	PC-15	0.9	PC-3	0.74	PC-3	0.73	PC-3	0.76
CLF16O	PC-11	0.88	PC-17	0.69	PC-20	0.56	PC-17	0.41
CPIAUC	PLS-3	0.9	RFSIRb^{c,d}-5.1	0.91	SIRb^{d}-4	0.97	R8SIRb^{c,d}-1	0.99
CES060	AR4	1	SIRb^{d}-1	0.99	R8SIRb^{c,d}-1	0.94	PC-1	0.92
DPCER	PC.ON^{f}-1.5	0.93	PC-1	0.92	RFSIRb^{c,d}-1	0.94	R8SIRb^{c,d}-1	0.94
RPI	PLS-1	0.91	PLS-1	0.85	PLS-1	0.82	PLS-1	0.8

Model Deninition

[b]The number after RIDGE Comp. is value of regularization param.

[c]R#SIR is regularized SIR. # refers to number of leading PCs used for regularization.

[d]In type–a SIR target is y_{t+h}. In type–b SIR target is y_t.

[e]RFSIR is regularized SIR on most frequently selected PCs by out-of-sample best subset selection.

[f]PC.ON is PCR in which number of components is chosen using Onatski criterion.

[g]PC.ICP1 is PCR in which number of components is chosen using Bai-Ng criterion.

[h]PC.BS is PCR in which number of components is chosen by best subset selection.

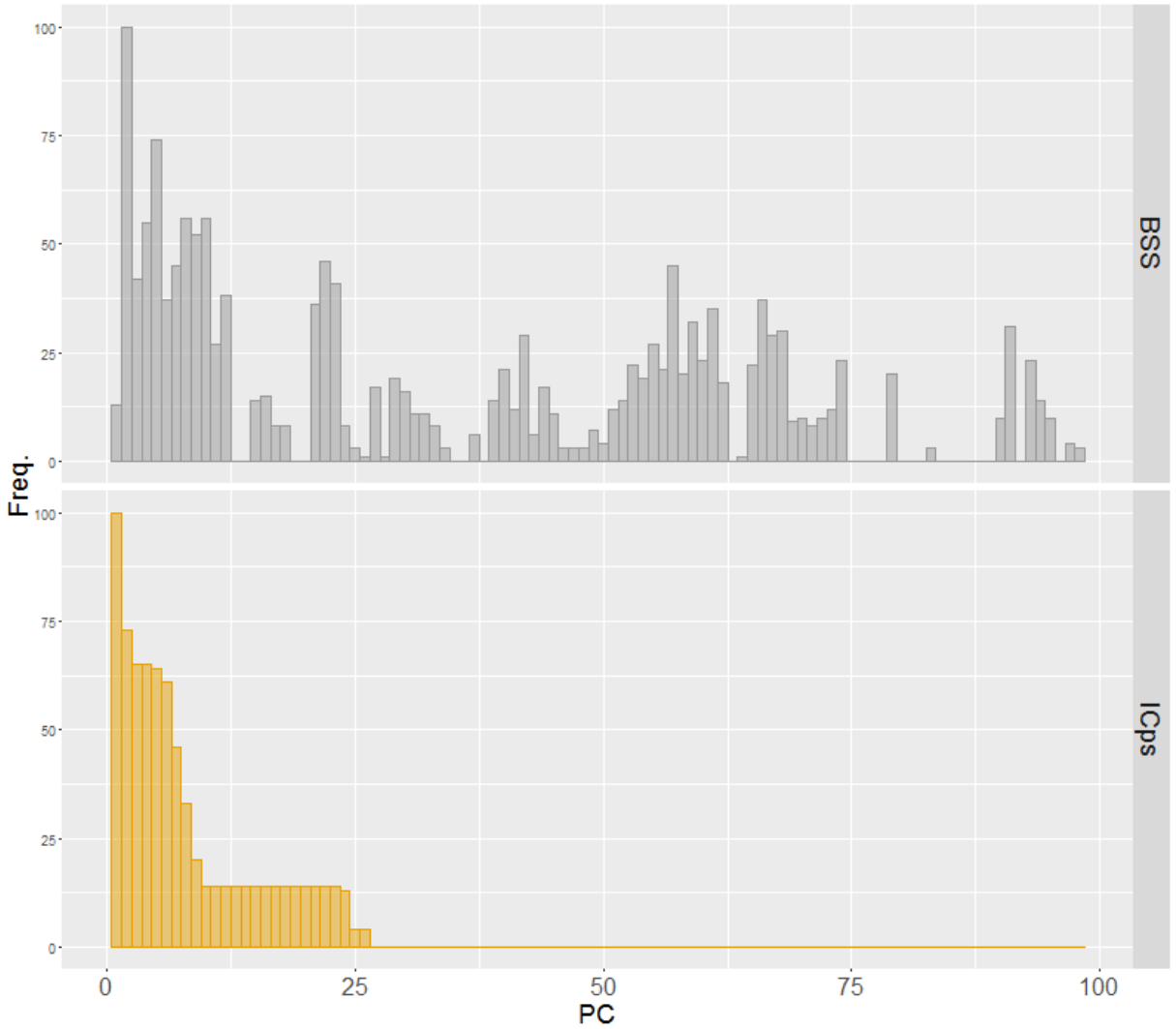
[i]PCF.BS is PCR in which number of components is chosen by best subset selection applied on most frequently chosen PCs in out-of-sample.

Near-Equivalence in Forecast Accuracy

- Great variability of best estimator depending on (target-sample-horizon) triplet.
- Sometimes beating AR(4) remains a challenging task.
- SIR–type estimators do fairly well, PLS is the estimator represented the least in the tables.
- Competitive edge of SIR is its parsimony: attains practically the same forecasting accuracy using one or two linear combinations of the predictors
- No Clear Winner result ubiquitous in macro-forecasting literature.
- Can we explain the results?

Selection of PCs Using Different Criteria

Figure: Av. Num. of PC Comp. Selected by BSS versus ICps Over 1992-2010.



MSFE As A Function of Number of Components

Common Forecasting Framework

Forecasting with Many Predictors

- Forecasting target variable y with a large set of predictors $\mathbf{x} \in \mathbb{R}^p$.
- Starting forecasting model includes all $\mathbf{x}$ predictors:

$$y_{t+h} = \alpha_0 + \boldsymbol{\alpha}_1' \mathbf{x}_t + \boldsymbol{\alpha}_2' \mathbf{w}_t + \varepsilon_{t+h}$$

where

  - $\mathbf{w}_t$ may contain additional regressors such as lags of y_t
  - $\mathrm{E}(\varepsilon_{t+h}) = 0$
  - $\boldsymbol{\alpha}_1' \mathbf{x}_t$ and $\boldsymbol{\alpha}_2' \mathbf{w}_t$ are uncorrelated with ε_{t+h}

- Naturally conducive to **OLS** to estimate $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ and the feasible forecast is

$$\widehat{y}_{t+h} = \hat{\alpha}_0 + \widehat{\boldsymbol{\alpha}}_1' \mathbf{x}_t + \widehat{\boldsymbol{\alpha}}_2' \mathbf{w}_t$$

Dimension Reduction Methods

- OLS estimation can be problematic when p is large relative to T, or variables in $\mathbf{x}_t$ are nearly collinear, as in macro forecasting $\Longrightarrow$ dimension reduction.

- Dimension reduction methods in regression fall into two categories:

  - **Variable Selection**: subset of the original predictors is selected for modeling the response (e.g. Stepwise Regression, All Subset Selection, LASSO) $\Longrightarrow$ Some predictors are discarded

  - **Feature Extraction**: linear combinations of the regressors replace the original regressors, therefore reduction of p happens prior to model fitting (e.g. PCR) $\Longrightarrow$ All predictors are retained.

Dimension Reduction Methods

- OLS estimation can be problematic when p is large relative to T, or variables in $\mathbf{x}_t$ are nearly collinear, as in macro forecasting $\Longrightarrow$ dimension reduction.

- Dimension reduction methods in regression fall into two categories:

  - **Variable Selection**: subset of the original predictors is selected for modeling the response (e.g. Stepwise Regression, All Subset Selection, LASSO) $\Longrightarrow$ Some predictors are discarded

  - **Feature Extraction**: linear combinations of the regressors replace the original regressors, therefore reduction of p happens prior to model fitting (e.g. PCR) $\Longrightarrow$ All predictors are retained.

Forecasting with Extracted Linear Features

Assuming

$$y_{t+h} = \boldsymbol{\alpha}' \mathbf{x}_t + \varepsilon_{t+h}$$

Forecasting with extracted features is implemented in two steps:

(i) Extract reduced features

$$\mathbf{f}_t = \boldsymbol{\beta}' \mathbf{x}_t$$

(ii) Fit the *reduced* model

$$\begin{aligned} y_{t+h} &= \boldsymbol{\gamma}' \mathbf{f}_t + \varepsilon_{t+h} \\ &= (\boldsymbol{\beta}\boldsymbol{\gamma})' \mathbf{x}_t + \varepsilon_{t+h} \end{aligned}$$

Justifying Reduced Model

Proposition

Suppose $\mathbf{x}$ *is a random p-vector with finite first two moments. Assume* $y = \boldsymbol{\alpha}'\mathbf{x} + \varepsilon$, *with* $\mathrm{E}(\varepsilon) = 0$ *and* $\mathrm{Cov}(\varepsilon, \mathbf{x}) = 0$ *where* $\boldsymbol{\alpha} \in \mathbb{R}^p$ *is unknown, and* $\mathbf{f} = \boldsymbol{\beta}'\mathbf{x}$, *where the* $p \times r$ *matrix* $\boldsymbol{\beta}$ *is such that*

(i) *for each* $\boldsymbol{\alpha}$, $\mathrm{E}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x} = \mathbf{f})$ *is linear in* $\mathbf{f} \in \mathbb{R}^r$ *(LC);*

(ii) *for each* $\boldsymbol{\alpha}$, $\mathrm{Var}(\boldsymbol{\alpha}'\mathbf{x}|\boldsymbol{\beta}'\mathbf{x} = \mathbf{f})$ *is constant in* $\mathbf{f} \in \mathbb{R}^r$ *(CVC).*

Then, y *can be decomposed into the sum of a linear function of* $\mathbf{f}$ *and a remainder or error term, as follows,*

$$y = \mu_y + \boldsymbol{\gamma}'(\mathbf{f} - \mathrm{E}(\mathbf{f})) + \epsilon$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^-\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\alpha} \in \mathbb{R}^r$, $\mu_y = \mathrm{E}(y)$, $\mathrm{E}(\epsilon|\mathbf{f}) = 0$ *and* $\mathrm{var}(\epsilon|\mathbf{f})$ *is constant.*

Discussion of Proposition

- It assumes the DGP for target is: $y_{t+h} = \boldsymbol{\alpha}' \mathbf{x}_t + \varepsilon_{t+h}$
- It places assumptions on marginal distribution of **observable** predictors
  - No assumptions of underlying latent factor structure plus assumptions on **unobservable** quantities.
- It concludes that the *reduced* forecasting model

$$y_{t+h} = \boldsymbol{\gamma}' \mathbf{f}_t + \varepsilon_{t+h}$$

  is "equivalent" to the DGP, with $\mathbf{f}_t$ the linear extracted features.
- We can remove (CVC) and lose homoskedasticity in regression error.

Do (LC) and (CVC) Hold in Practice?

Conditions (LC) and (CVC) are difficult to verify in practice since β is unknown. However:

- Conditions (LC) and (CVC) are satisfied for any β if $\mathbf{x}$ is multivariate normal.
  - Joint normality on FRED-MD is rejected.
- Condition (LC) is satisfied for all β if $\mathbf{x}$ has an **elliptically contoured** distribution (Eaton (1986))
  - *FRED-MD satisfies ellipticity tests*.
- Therefore with (LC) satisfied, we have first piece of the puzzle why different feature extraction methods have similar forecast accuracy.

Feature Extraction

- Feature extraction is carried out by solving:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i=1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \text{Corr}^2(h(y), \mathbf{x}'\boldsymbol{\beta}_i) \times H[\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)]$$

where

  - OLS: $H(\cdot) = 1$, $h(\cdot) = \textit{identity}$
  - PCR: $H(\cdot) = \textit{identity}$, $h(\cdot) = 1$
  - PLS: $H(\cdot) = \textit{identity}$, $h(\cdot) = \textit{identity}$
  - RIDGE: $H(\cdot) = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)+\kappa}$, $h(\cdot) = \textit{identity}$
  - SIR: $H(\cdot) = \textit{identity}$, $h(\cdot) = \text{E}[\boldsymbol{\beta}_i'\mathbf{x}|y]$

Feature Extraction

- Feature extraction is carried out by solving:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i=1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \text{Corr}^2(h(y), \mathbf{x}'\boldsymbol{\beta}_i) \times H[\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)]$$

where

  - OLS: $H(\cdot) = 1$, $h(\cdot) = identity$
  - PCR: $H(\cdot) = identity$, $h(\cdot) = 1$
  - PLS: $H(\cdot) = identity$, $h(\cdot) = identity$
  - RIDGE: $H(\cdot) = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)+\kappa}$, $h(\cdot) = identity$
  - SIR: $H(\cdot) = identity$, $h(\cdot) = \text{E}[\boldsymbol{\beta}_i'\mathbf{x}|y]$

Feature Extraction

- Feature extraction is carried out by solving:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i=1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \text{Corr}^2(h(y), \mathbf{x}'\boldsymbol{\beta}_i) \times H[\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)]$$

where

  - OLS: $H(\cdot) = 1$, $h(\cdot) = identity$
  - PCR: $H(\cdot) = identity$, $h(\cdot) = 1$
  - PLS: $H(\cdot) = identity$, $h(\cdot) = identity$
  - RIDGE: $H(\cdot) = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)+\kappa}$, $h(\cdot) = identity$
  - SIR: $H(\cdot) = identity$, $h(\cdot) = \text{E}[\boldsymbol{\beta}_i'\mathbf{x}|y]$

Feature Extraction

- Feature extraction is carried out by solving:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i=1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \text{Corr}^2(h(y), \mathbf{x}'\boldsymbol{\beta}_i) \times H[\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)]$$

where

  - OLS: $H(\cdot) = 1$, $h(\cdot) = identity$
  - PCR: $H(\cdot) = identity$, $h(\cdot) = 1$
  - PLS: $H(\cdot) = identity$, $h(\cdot) = identity$
  - RIDGE: $H(\cdot) = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}_i)+\kappa}$, $h(\cdot) = identity$
  - SIR: $H(\cdot) = identity$, $h(\cdot) = \text{E}[\boldsymbol{\beta}_i'\mathbf{x}|y]$

Feature Extraction

- Feature extraction is carried out by solving:

$$\max_{\substack{\{\boldsymbol{\beta}_i\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\beta}_i=1\} \\ \{\boldsymbol{\beta}_i'\boldsymbol{\Sigma}_x\boldsymbol{\beta}_j=0\}_{j=1}^{i-1}}} \quad \mathrm{Corr}^2(h(y), \mathbf{x}'\boldsymbol{\beta}_i) \times H[\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}_i)]$$

where

  - OLS: $H(\cdot) = 1$, $h(\cdot) = identity$
  - PCR: $H(\cdot) = identity$, $h(\cdot) = 1$
  - PLS: $H(\cdot) = identity$, $h(\cdot) = identity$
  - RIDGE: $H(\cdot) = \frac{\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}_i)}{\mathrm{Var}(\mathbf{x}'\boldsymbol{\beta}_i)+\kappa}$, $h(\cdot) = identity$
  - SIR: $H(\cdot) = identity$, $h(\cdot) = \mathrm{E}[\boldsymbol{\beta}_i'\mathbf{x}|y]$

Prediction

- Estimate *reduced* predictors: $\widehat{\mathbf{f}}_t = \widehat{\boldsymbol{\beta}}' \mathbf{x}_t$
- Fit the reduced model: $y_{t+h} = \boldsymbol{\gamma}'(\widehat{\boldsymbol{\beta}}' \mathbf{x}_t) + \varepsilon_{t+h}$
- Prediction coefficient: $\mathbf{b} = \widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\gamma}}$
- Prediction is: $\widehat{y}_{t+h|t} = \mathbf{b}'\mathbf{x}_0 = \widehat{\boldsymbol{\gamma}}'\widehat{\boldsymbol{\beta}}'\mathbf{x}_0$
- Common pattern for $\mathbf{b}$ across feature extraction methods:

$$\mathbf{b} = \text{scaling factor} * \text{signal}$$

.

Table: Summary of Estimators for $y = \boldsymbol{\alpha}'\mathbf{x} + \epsilon$

	Metaparameter	*Scaling Factor*	*Signal*
OLS	–	$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$	$\boldsymbol{\sigma}_{\mathbf{x}y}$
RIDGE	κ	$(\boldsymbol{\Sigma}_{\mathbf{x}} + \kappa\mathbf{I})^{-1}$	$\boldsymbol{\sigma}_{\mathbf{x}y}$
PCR	m	$\boldsymbol{\Sigma}_{\mathbf{x}}^{-}(m)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$
PLS	s	$\mathbf{W}_{\mathrm{PLS}}(s)\left(\mathbf{W}'_{\mathrm{PLS}}(s)\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{W}_{\mathrm{PLS}}(s)\right)^{-1}\mathbf{W}'_{\mathrm{PLS}}(s)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$
SIR	d	$\mathbf{W}_{\mathrm{SIR}}(d)\left(\mathbf{W}'_{\mathrm{SIR}}(d)\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{W}_{\mathrm{SIR}}(d)\right)^{-1}\mathbf{W}'_{\mathrm{SIR}}(d)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$
RSIR	m, d	$\mathbf{W}_{\mathrm{RSIR}}(d)\left(\mathbf{W}'_{\mathrm{RSIR}}(d)\boldsymbol{\Sigma}_{\mathbf{x}}(m)\mathbf{W}_{\mathrm{RSIR}}(d)\right)^{-1}\mathbf{W}'_{\mathrm{RSIR}}(d)$	$\boldsymbol{\sigma}_{\mathbf{x}y}$

These estimators have

- Same **signal** ($\boldsymbol{\sigma}_{\mathbf{x},y}$): *near-equivalent forecast accuracy*

- Different **scaling factor**: *dimension*

Focus on Targeted Estimators

- For **targeted** estimators PLS and SIR scaling factor uses y.
- How they differ:
  - **PLS**: $\mathbf{W}_{\mathrm{PLS}}(s) = \left(\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{s-1}\boldsymbol{\sigma}_{\mathbf{x}y}\right)$
  - **SIR**: $\mathbf{W}_{\mathrm{SIR}}(d) = \left(\mathrm{E}(\mathbf{x}|y), \boldsymbol{\Sigma}_{\mathbf{x}}\,\mathrm{E}(\mathbf{x}|y), \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{d-1}\,\mathrm{E}(\mathbf{x}|y)\right)$
- Therefore the reduction is based:
  - For **PLS** on a moment of the **joint** distribution $F(y, \mathbf{x})$.
  - For **SIR** on a moment of the **conditional** distribution $F(y|\mathbf{x})$.

Focus on Targeted Estimators

- For **targeted** estimators PLS and SIR scaling factor uses y.
- How they differ:
  - **PLS**: $\mathbf{W}_{\mathrm{PLS}}(s) = \left(\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{s-1}\boldsymbol{\sigma}_{\mathbf{x}y}\right)$
  - **SIR**: $\mathbf{W}_{\mathrm{SIR}}(d) = \left(\mathrm{E}(\mathbf{x}|y), \boldsymbol{\Sigma}_{\mathbf{x}}\,\mathrm{E}(\mathbf{x}|y), \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{d-1}\,\mathrm{E}(\mathbf{x}|y)\right)$
- Therefore the reduction is based:
  - For **PLS** on a moment of the **joint** distribution $F(y, \mathbf{x})$.
  - For **SIR** on a moment of the **conditional** distribution $F(y|\mathbf{x})$.

Focus on Targeted Estimators

- For **targeted** estimators PLS and SIR scaling factor uses y.
- How they differ:
  - **PLS**: $\mathbf{W}_{\mathrm{PLS}}(s) = \left(\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{s-1}\boldsymbol{\sigma}_{\mathbf{x}y}\right)$
  - **SIR**: $\mathbf{W}_{\mathrm{SIR}}(d) = \left(\mathrm{E}(\mathbf{x}|y), \boldsymbol{\Sigma}_{\mathbf{x}}\,\mathrm{E}(\mathbf{x}|y), \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{d-1}\,\mathrm{E}(\mathbf{x}|y)\right)$
- Therefore the reduction is based:
  - For **PLS** on a moment of the **joint** distribution $F(y, \mathbf{x})$.
  - For **SIR** on a moment of the **conditional** distribution $F(y|\mathbf{x})$.

Focus on Targeted Estimators

- For **targeted** estimators PLS and SIR scaling factor uses y.
- How they differ:
  - **PLS**: $\mathbf{W}_{\mathrm{PLS}}(s) = \left(\boldsymbol{\sigma}_{\mathbf{x}y}, \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\sigma}_{\mathbf{x}y}, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{s-1}\boldsymbol{\sigma}_{\mathbf{x}y}\right)$
  - **SIR**: $\mathbf{W}_{\mathrm{SIR}}(d) = \left(\mathrm{E}(\mathbf{x}|y), \boldsymbol{\Sigma}_{\mathbf{x}}\,\mathrm{E}(\mathbf{x}|y), \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}{}^{d-1}\,\mathrm{E}(\mathbf{x}|y)\right)$
- Therefore the reduction is based:
  - For **PLS** on a moment of the **joint** distribution $F(y, \mathbf{x})$.
  - For **SIR** on a moment of the **conditional** distribution $F(y|\mathbf{x})$.

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Why is SIR more effective?

$$\text{var}(\mathbf{X}) = \text{var}[\text{E}(\mathbf{X}|Y)] + \text{E}[\text{var}(\mathbf{X}|Y)]$$

- For simplicity, assume Y is categorical: $\mathbf{X}|Y$ is the restriction of $\mathbf{X}$ in the class defined by Y
- **Signal:** $\text{var}[\text{E}(\mathbf{X}|Y)]$ is between group variation in $\mathbf{X}$
- **Noise:** $\text{E}[\text{var}(\mathbf{X}|Y)]$ is within group variation
- **PCR** mixes up noise and signal when extracting PCs
- **PLS** produces ordering of eigen-components according to their importance to $\text{cov}(\mathbf{X}, Y)$, i.e. captures linear dependence of $\mathbf{X}$ and Y
- **SIR** produces ordering of eigen-components according to their importance to Y, linear and non-linear

Beyond Linear Signals: Sufficient Dimension Reduction

Sufficient Reductions

- Suppose $\mathbf{R}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p$.
- $\mathbf{R}(\mathbf{x})$ is a **Sufficient Reduction** when no information about y is lost when $\mathbf{x}$ is replaced by $\mathbf{R}(\mathbf{x})$, i.e.

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x}))$$

- How to use in forecasting: For example, if the reduction is linear,

$$\mathbf{R}(\mathbf{x}) = \beta'\mathbf{x}$$

the forecasting model is

$$y_{t+h} = g(\beta'\mathbf{x}, \varepsilon_{t+h})$$

Sufficient Reductions

- Suppose $\mathbf{R}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p$.
- $\mathbf{R}(\mathbf{x})$ is a **Sufficient Reduction** when no information about y is lost when $\mathbf{x}$ is replaced by $\mathbf{R}(\mathbf{x})$, i.e.

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x}))$$

- How to use in forecasting: For example, if the reduction is linear,

$$\mathbf{R}(\mathbf{x}) = \beta'\mathbf{x}$$

the forecasting model is

$$y_{t+h} = g(\beta'\mathbf{x}, \varepsilon_{t+h})$$

Sufficient Reductions

- Suppose $\mathbf{R}(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^d$ with $d \leq p$.
- $\mathbf{R}(\mathbf{x})$ is a **Sufficient Reduction** when no information about y is lost when $\mathbf{x}$ is replaced by $\mathbf{R}(\mathbf{x})$, i.e.

$$F(y|\mathbf{x}) = F(y|\mathbf{R}(\mathbf{x}))$$

- How to use in forecasting: For example, if the reduction is linear,

$$\mathbf{R}(\mathbf{x}) = \beta'\mathbf{x}$$

the forecasting model is

$$y_{t+h} = g(\beta'\mathbf{x}, \varepsilon_{t+h})$$

Identifying Sufficient Linear Reductions

Linear SDR allows identification and estimation of $\mathcal{C}(\beta)$ with $F(Y|\mathbf{X}) = F(Y|\beta^T\mathbf{X})$

- General Idea: find a kernel matrix $\mathbf{M}$ so that

$$\text{span}(\mathbf{M}) \subset \text{span}(\beta)$$

- SDR methods: different proposals for $\mathbf{M}$

For example: In SIR

If $\mathrm{E}(\mathbf{x}|\beta'\mathbf{x})$ is linear in $\beta'\mathbf{x}$ for any β (LC) such that $F(y|\mathbf{x}) = F(y|\beta'\mathbf{x})$, then

$$\begin{aligned}\mathcal{C}\left(\mathbf{\Sigma}_{\mathbf{x}}^{-1}[\mathrm{E}(\mathbf{x}|y) - \mathrm{E}(\mathbf{x})]\right) &= \mathcal{C}\left(\mathrm{Var}[\mathrm{E}(\mathbf{x}|y)]\right) \\ &\subseteq \mathcal{C}(\beta)\end{aligned}$$

In SIR,

$$\mathbf{M} = \mathbf{\Sigma}_x^{-1}\,\mathrm{Var}(\mathrm{E}(\mathbf{X}|\,Y))$$

Forecasting with Sliced Inverse Regression

- SIR Reduces the predictors:
  - (i) Computes non-parametric estimate of $\mathrm{Var}(\mathrm{E}(\mathbf{x}|y))$
  - (ii) Sets $\widehat{\beta}_d(SIR)$ to be the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}\widehat{\mathrm{Var}(\mathrm{E}(\mathbf{x}|y))}$
- Run non-parametric regression $y_{t+h} = g(\widehat{\beta}_d(SIR)'\mathbf{x}_t, \varepsilon_{t+h})$ and predict.
  - We did not find any forecasting improvement with non-linear $g(\cdot)$.
- RSIR formed by substituting $\mathbf{x}$ with PCs of $\mathbf{\Sigma}_x$.

Conclusions

Comparing Forecasting Frameworks

Forecasting Framework	*DGP for* y_{t+h}	*DGP for* $\mathbf{x}_t$
	DFM	
Standard	$y_{t+h} = \boldsymbol{\gamma}'\mathbf{f}_t + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}\mathbf{f}_t + \mathbf{u}_t$
With Irrelevant Factors	$y_{t+h} = \boldsymbol{\gamma}'\mathbf{f}_{1,t} + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}_1\mathbf{f}_{1,t} + \boldsymbol{\Gamma}_2\mathbf{f}_{2,t} + \mathbf{u}_t$
No Factors for Target	$y_{t+h} = \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}\mathbf{f}_t + \mathbf{u}_t$
	Other	
No Factors	$y_{t+h} = \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_{t+h}$	Conditions on $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\beta}$
	SDR	
Non-linear Forecast	$y_{t+h} = g(\boldsymbol{\alpha}'\mathbf{x}_t, \varepsilon_{t+h})$	$\mathrm{E}[\mathbf{x}_t \vert \boldsymbol{\beta}'\mathbf{x}_t] = \mathbf{A}\boldsymbol{\beta}'\mathbf{x}_t$
Linear Forecast	$y_{t+h} = \boldsymbol{\gamma}'(\boldsymbol{\alpha}'\mathbf{x}_t) + \varepsilon_{t+h}$	$\mathrm{E}[\mathbf{x}_t \vert \boldsymbol{\beta}'\mathbf{x}_t] = \mathbf{A}\boldsymbol{\beta}'\mathbf{x}_t$

- Most forecasting literature studies shrinkage under factor structure.
- We study shrinkage estimators under different DGP for observables.

Comparing Forecasting Frameworks

Forecasting Framework	*DGP for* y_{t+h}	*DGP for* $\mathbf{x}_t$
	DFM	
Standard	$y_{t+h} = \boldsymbol{\gamma}'\mathbf{f}_t + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}\mathbf{f}_t + \mathbf{u}_t$
With Irrelevant Factors	$y_{t+h} = \boldsymbol{\gamma}'\mathbf{f}_{1,t} + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}_1\mathbf{f}_{1,t} + \boldsymbol{\Gamma}_2\mathbf{f}_{2,t} + \mathbf{u}_t$
No Factors for Target	$y_{t+h} = \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_{t+h}$	$\mathbf{x}_t = \boldsymbol{\Gamma}\mathbf{f}_t + \mathbf{u}_t$
	Other	
No Factors	$y_{t+h} = \boldsymbol{\alpha}'\mathbf{x}_t + \varepsilon_{t+h}$	Conditions on $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\beta}$
	SDR	
Non-linear Forecast	$y_{t+h} = g(\boldsymbol{\alpha}'\mathbf{x}_t, \varepsilon_{t+h})$	$\mathrm{E}[\mathbf{x}_t \vert \boldsymbol{\beta}'\mathbf{x}_t] = \mathbf{A}\boldsymbol{\beta}'\mathbf{x}_t$
Linear Forecast	$y_{t+h} = \boldsymbol{\gamma}'(\boldsymbol{\alpha}'\mathbf{x}_t) + \varepsilon_{t+h}$	$\mathrm{E}[\mathbf{x}_t \vert \boldsymbol{\beta}'\mathbf{x}_t] = \mathbf{A}\boldsymbol{\beta}'\mathbf{x}_t$

- Most forecasting literature studies shrinkage under factor structure.
- We study shrinkage estimators under different DGP for observables.

Data Specific Conclusions

- Since there is no non-linear trend in $\mathrm{E}(y|\mathbf{x})$ and $\mathbf{x}$ is elliptically contoured, Proposition 1 yields that the reduced model

$$y_{t+h} = \boldsymbol{\gamma}'(\boldsymbol{\beta}'\mathbf{x}_t) + \epsilon_{t+h}$$

  is a good approximation model to the true population model.

- What does this mean for prediction?

$$\mathrm{E}(y-\mu_y|\boldsymbol{\beta}'\mathbf{x}) = \boldsymbol{\alpha}'\boldsymbol{\Sigma}_x\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^-\boldsymbol{\beta}'(\mathbf{x}-\mathrm{E}(\mathbf{x})) = \boldsymbol{\alpha}'\mathbf{P}'_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)}(\mathbf{x}-\mathrm{E}(\mathbf{x}))$$

where $\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)} = \boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}_x\boldsymbol{\beta})^-\boldsymbol{\beta}'\boldsymbol{\Sigma}_x$ is the projection operator onto $\mathrm{span}(\boldsymbol{\beta})$ relative to $(\mathbf{a},\mathbf{b}) = \mathbf{a}'\boldsymbol{\Sigma}_x\mathbf{b}$.

- For a given $\mathbf{x}$ value, how close $\mathrm{E}(y|\mathbf{x})$ will be to the truth is reflected by the norm of $\mathbf{I}-\mathbf{P}_{\boldsymbol{\beta}(\boldsymbol{\Sigma}_x)}$, which is controlled solely by $\boldsymbol{\beta}$.

- In consequence, *ordering estimators $\boldsymbol{\beta}'\mathbf{x}$ with respect to their forecasting accuracy is tantamount to identifying $\boldsymbol{\beta}$'s with smaller norm.*

Can we improve the forecasting ability of the model?

- Forecasting accuracy results do not support non-linearities in the **conditional mean** of the target variable
  - If there is non-linear signal, it must be in the variance
- Dimension two or higher in SIR indicates the presence of nonlinear relationships
  - Plots of the residuals of the fitted forecasting model versus the second SIR component indicate that the variance of the forecasting models varies for most targets

Can we improve the forecasting ability of the model?

- The data passed the multivariate elliptical symmetry test but not the normality test
  - (LC) holds but not (NCV)
  - Model-based SDR: (Bura and Forzani (2015))
    - When $\mathbf{X}|Y \sim EC_p(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}, g_Y)$, then

$$\mathbf{R}(\mathbf{X}) = (\boldsymbol{\alpha}'(\mathbf{X} - \mathrm{E}(\mathbf{X})), (\mathbf{X} - \mathrm{E}(\mathbf{X}))'\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$$

    - The **minimal sufficient reduction has a non-linear component**
  - Need to model the variance using $(\mathbf{X} - \mathrm{E}(\mathbf{X}))'\boldsymbol{\Sigma}_x^{-1}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$

References

- Bai J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, **146**(2), 304-317.

- Barbarino, A. and Bura, E. (2015). Forecasting with sufficient dimension reductions. Finance and Economics Discussion Series 2015-074. Washington: Board of Governors of the Federal Reserve System, http://dx.doi.org/10.17016/FEDS.2015.074.

- Bura, E. and Cook, R. D. (2001). Extending SIR: The Weighted Chi-Square Test, *Journal of the American Statistical Association* **96**, 996–1003.

- Bura, E. and Yang, J. (2011). Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach. *Journal of Multivariate Analysis*, **102**, 130-142.

- Bura, E. and Forzani, J. (2014). Sufficient reductions in regressions with elliptically contoured inverse predictors. *Journal of the American Statistical Association* 110, 420-434

- Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2005), The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting, *Journal of the American Statistical Association* 100(471), 830-840.

- Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.
- McCracken, M.W. and Ng, S. (2015), FRED-MD: A Monthly Database for Macroeconomic Research, Working Paper 2015-012A, June 2015.
- Steinberger, L. and Leeb, H. (2015), On conditional moments of high-dimensional random vectors given lower-dimensional projections, `http://arxiv.org/pdf/1405.2183.pdf`.
- Stock, J. H. and Watson, M. 2002a. Macroeconomic forecasting using diffusion indices. *Journal of Business and Economics Statistics*, **20**, 147162.
- Stock, J. H. and Watson, M. 2002b. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 11671179.
- Stock, J. H. and Watson, M. (2006), Macroeconomic Forecasting Using Many Predictors, *Handbook of Economic Forecasting*, Graham Elliott, Clive Granger, Allan Timmerman (eds.), North Holland.