CLUSTER-BASED REGULARIZED SLICED INVERSE REGRESSION FOR FORECASTING MACROECONOMIC VARIABLES*

YU Yue · CHEN Zhihong · YANG Jie

DOI: 10.1007/s11424-014-3281-8
Received: 13 April 2012 / Revised: 1 May 2013

Abstract This paper concerns the dimension reduction in regression for large data set. The authors introduce a new method based on the sliced inverse regression approach, called cluster-based regularized sliced inverse regression. The proposed method not only keeps the merit of considering both response and predictors' information, but also enhances the capability of handling highly correlated variables. It is justified under certain linearity conditions. An empirical application on a macroeconomic data set shows that the proposed method has outperformed the dynamic factor model and other shrinkage methods.

Keywords Cluster-based, forecast, macroeconomics, sliced inverse regression.

1 Introduction

Forecasting using many predictors has received a good deal of attention in recent years. The curse of dimensionality has been turned into a blessing with the abundant information in large datasets. Various methods have been originated to extract efficient predictors, for example, dynamic factor model (DFM), Bayesian model averaging, Lasso, boosting, etc. Among them, dynamic factor model is conceptually appealing in macroeconomics because it is structurally consistent with log-linearlized models such as dynamic stochastic general equilibrium models.

YU Yue
TradeLink L.L.C., 71 *S. Wacker Drive Suite* 1900*, Chicago, Illinois, USA.* Email: yuyue@trdlnk.com.
CHEN Zhihong
School of International Trade and Economics, University of International Business and Economics, Beijing 100029*, China.* Email: zhihong.chen@uibe.edu.cn.
YANG Jie
Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois, USA. Email: jyang06@math.uic.edu.

*This paper was supported by the National Science Foundation of China under Grant No. 71101030 and the Program for Innovative Research Team in UIBE under Grant No. CXTD4-01.
◇ *This paper was recommended for publication by Editor WANG Shouyang.*

Boivin and Ng[1] assessed the extent to which the forecasts are influenced by how the factors are estimated and/or how the forecasts are formulated. They did not find one method that always stands out to be systematically good or bad. Meta-study from Eickmeier and Ziegler[2] also found mixed performance of DFM forecasts. Stock and Watson[3] compared the dynamic factor model with some recent multi-predictor methods. They concluded that the dynamic factor model could not be outperformed by these methods for all the forecasting series in their data set.

The recent development in statistics provides a new method of dimension reduction in regression for large-dimensioned data. The literature stems from Duan and Li[4], and Li[5], which proposed a new way of thinking in the regression analysis, called sliced inverse regression (SIR). SIR reverses the role of response y and predictors $\boldsymbol{x}$. Classical regression methods mainly deal with the conditional density $\boldsymbol{f}(y|\boldsymbol{x})$. SIR collects the information of the variation of predictors $\boldsymbol{x}$ along with the change of the response y, by exploring the conditional density $h(\boldsymbol{x}|y)$. Usually the dimension of the response is far more less than the dimension of the predictors, hence, it is a way to avoid the "curse of dimensionality".

The traditional SIR does not work well for highly correlated data, due to the degenerate covariance matrix. This is not feasible when the number of variables N is greater than the number of observations T, which happens a lot in economics studies. In addition, the economic variables are often highly correlated or inversely correlated, due to the derivation formula, data sources, and grouping category, for instance, personal consumption expenditures (PCE) and consumer price index (CPI), total employees, and unemployment rate, etc. This makes the covariance matrix ill-conditioned, causes the inverse matrix lack of precision and too sensitive to the variation of matrix entries, and leads to unstable results with large standard deviations. There are some extensions of SIR for the highly collinearity data and "$T < N$" problems, for example, regularized sliced inverse regression (Zhong, et al.[6], Li and Yin[7]), and partial inverse regression (Li, et al.[8]).

In this article, we propose a new method of dimension reduction, called the cluster-based sliced inverse regression (CRSIR) method, for many predictors in a data rich environment. We evaluate its properties theoretically and use it for forecasting macroeconomic series. Comparison in terms of both in-sample prediction and out-of-sample forecasting simulation shows the advantage of our method.

The remaining of the article is organized as follows. Section 2 introduces cluster-based SIR method with its statistical property. An empirical application on the macroeconomic dataset used by Stock and Watson[3] is given in Section 3. Conclusions with some discussions are provided in Section 4.

2 Modeling and Methods

The regression model in [5] takes the form of

$$y = g(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{\beta}_2^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K^{\mathrm{T}}\boldsymbol{x}, \varepsilon), \tag{1}$$

where the response y is univariate, $\boldsymbol{x}$ is an N-dimensional vector, and the random error $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{x}$. Figure 1 gives a straightforward illustration of Model (1), which means that y depends on $\boldsymbol{x}$ only through the K-dimensional subspace spanned by projection vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_K$, known as the effective dimension reducing directions (e.d.r.-directions)[5].

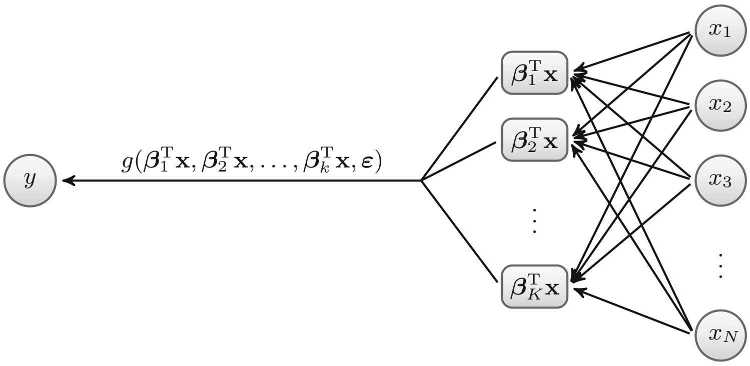

y
$g(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}, \ldots, \boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{x}, \boldsymbol{\varepsilon})$
$\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}$
$\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x}$
$\vdots$
$\boldsymbol{\beta}_K^{\mathrm{T}}\mathbf{x}$
x_1
x_2
x_3
$\vdots$
x_N

Figure 1 Regression Model (1) using e.d.r.-directions

Many methods can be used to find the e.d.r.-directions, for example, principal component analysis might be the most commonly used one in economics. But unlike these methods, SIR not only reduces dimensions in regression but also integrates the information from both predictors and response. Moreover, different from the classical regression methods, SIR intends to collect information on how $\boldsymbol{x}$ changes along with y. That is to say, instead of estimating the forward regression function $\boldsymbol{\eta}(\boldsymbol{x}) = E(y|\boldsymbol{x})$, inverse regression considers $\boldsymbol{\xi}(y) = E(\boldsymbol{x}|y)$. Compared with $\boldsymbol{\eta}(\boldsymbol{x})$, the inverse regression function $\boldsymbol{\xi}(y)$ depends on one-dimensioned y, which makes the operation much easier.

Li[5] showed that using SIR method, the e.d.r.-directions can be estimated by solving

$$\mathrm{Cov}\big(E(\boldsymbol{x}|y)\big)\boldsymbol{\beta}_j = \nu_j \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_j, \tag{2}$$

where ν_j is the jth eigenvalue and $\boldsymbol{\beta}_j$ is the corresponding eigenvector of $\mathrm{Cov}\big(E(\boldsymbol{x}|y)\big)$ with respect to $\mathrm{Cov}(\boldsymbol{x})$. During the forecasting procedure, the covariance matrices can be replaced by their usual moment estimates.

One of the key parameters used in SIR is the number of slices H. However, it is not crucial for larger sample sizes, Li[9] indicated that for a sample size $n = 300$, H can be chosen between 10 to 20, and SIR outputs do not change much for a wide range of H. Accordingly, we fix $H = 10$ throughout this article.

2.1 Cluster-Based Sliced Inverse Regression

In this section, we introduce clustering methodology with the sliced inverse regression to improve the performance of SIR on collinear data.

Assuming that the variables of interest can be clustered into several blocks, so that two variables within the same block are correlated to each other, and any two variables belonging to different blocks are independent. In practice, an orthogonalization procedure can be applied to reduce the correlations between blocks in order to fit our assumption. Thus, we can cluster the variables according to their correlations in order to find the e.d.r-directions, because there is no shared information between clusters.

The clustering method we use is hierarchical clustering[10] with complete linkage. The dissimilarity is defined as $1 - |\text{Correlation}|$.

The algorithm for the cluster-based SIR method can be described as following:

1) Standardize each explanatory variable to zero mean and unit variance.

2) Cluster $\boldsymbol{x}$ $(N \times 1)$ into $\begin{pmatrix} \boldsymbol{x}_1^{\rm T} & \boldsymbol{x}_2^{\rm T} & \cdots & \boldsymbol{x}_c^{\rm T} \end{pmatrix}^{\rm T}$ based on the correlation matrix of $\boldsymbol{x}$, where $\boldsymbol{x}_i$ is $N_i \times 1$, $\sum_{i=1}^{c} N_i = N$, and c is the number of clusters, which will be determined by cross-validation.

3) Restricted to each cluster, perform SIR method and pick up k_i SIR directions based on the sequential chi-square test[5], say $\boldsymbol{\theta}_j^{(i)},\ j = 1, 2, \cdots, k_i,\ i = 1, 2, \cdots, c$.

4) Collect all the SIR variates obtained from the clusters, say $\{\boldsymbol{\theta}_j^{(i)\rm T}\boldsymbol{x}_i, i = 1, 2, \cdots, c, j = 1, 2, \cdots, k_i\}$.

5) Let $\boldsymbol{\lambda}_l = \begin{pmatrix} \mathbf{0_1}^{\rm T} & \boldsymbol{\theta}_j^{(i)\rm T} & \mathbf{0_2}^{\rm T} \end{pmatrix}^{\rm T}$, $l = 1, 2, \cdots, m$, $m = \sum_{i=1}^{c} k_i$, where $\mathbf{0_1}$ and $\mathbf{0_2}$ are zero column vectors with dimension $\sum_{k=1}^{i-1} N_k$ and $\sum_{k=i+1}^{c} N_k$, respectively. Denote $\mathit{\Lambda} = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \cdots, \boldsymbol{\lambda}_m)$. The variates $\{\boldsymbol{\theta}_j^{(i)\rm T}\boldsymbol{x}_i\}$ can be written in a vector form as $(\boldsymbol{\lambda}_1^{\rm T}\boldsymbol{x}, \boldsymbol{\lambda}_2^{\rm T}\boldsymbol{x}, \cdots, \boldsymbol{\lambda}_m^{\rm T}\boldsymbol{x})^{\rm T} = \mathit{\Lambda}^{\rm T}\boldsymbol{x}$.

6) Then, perform SIR method one more time to the pooled variates $\mathit{\Lambda}^{\rm T}\boldsymbol{x}$ to reduce dimensions further, and get the e.d.r.-directions $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_v)$, where v is also determined by the sequential chi-square test. Denote $\mathit{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_v)$, the final CRSIR variates we chose are $\mathit{\Gamma}^{\rm T}\mathit{\Lambda}^{\rm T}\boldsymbol{x}$.

7) Estimate the values of forecasting series using the CRSIR variates $\mathit{\Gamma}^{\rm T}\mathit{\Lambda}^{\rm T}\boldsymbol{x}$. Linear model with ordinary least squares (OLS) is used in this article, and as to be shown later, it is sufficiently good for our method.

Note that the matrices $\mathit{\Gamma}$ is $m \times v$, $\mathit{\Lambda}$ is $N \times m$, so $\mathit{\Gamma}^{\rm T}\mathit{\Lambda}^{\rm T}\boldsymbol{x}$ is $v \times 1$. Therefore, we only use v factors to build the final model for forecasting y, instead of using N variables based on the original dataset.

2.2 Statistical Property of Cluster-Based SIR

Li[5] established the unbiasedness for the e.d.r.-directions found by SIR, assuming the following linearity condition.

Linearity Condition 2.1 For any $\boldsymbol{b} \in \mathbb{R}^N$, the conditional expectation $E(\boldsymbol{b}^{\rm T}\boldsymbol{x}|\boldsymbol{\beta}_1^{\rm T}\boldsymbol{x}, \boldsymbol{\beta}_2^{\rm T}\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K^{\rm T}\boldsymbol{x})$ is linear in $\boldsymbol{\beta}_1^{\rm T}\boldsymbol{x}, \boldsymbol{\beta}_2^{\rm T}\boldsymbol{x}, \cdots, \boldsymbol{\beta}_K^{\rm T}\boldsymbol{x}$.

The linearity condition is not easy to verify, however, Eaton[11] showed when $\boldsymbol{x}$ is elliptically symmetrically distributed, for example, multivariate normally distributed, the linearity condition holds. Furthermore, Hall and Li[12] showed that elliptical symmetric distribution is not a restrictive assumption, because the linearity condition holds approximately when N is large even if the dataset has not been generated from an elliptically symmetric distribution.

Without loss of generality, we assume each variable in $\boldsymbol{x}$ has been standardized to zero mean and unit variance for our discussion. Li[5] proved the following theorem.

Theorem 2.2 *Assume Linearity Condition* 2.1*, the centered inverse regression curve* $E(\boldsymbol{x}|y)$ *is contained in the space spanned by* $\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}_j,\ j=1,2,\cdots,K$*, where* $\Sigma_{\boldsymbol{x}}$ *is the covariance matrix of* $\boldsymbol{x}$*.*

Figure 2 shows a three-dimensional case when $\boldsymbol{x}=\left(x_1,x_2,x_3\right)^{\mathrm{T}}$, since the inverse regression function $E(\boldsymbol{x}|y)$ is a function of y, it draws a curve in the three-dimensional space when y changes. Theorem 2.2 indicates that such curve is located exactly on the plane spanned by two directions d_1 and d_2 from $\Sigma_{\boldsymbol{x}}\boldsymbol{\beta}_j,\ j=1,2$, assuming $K=2$.

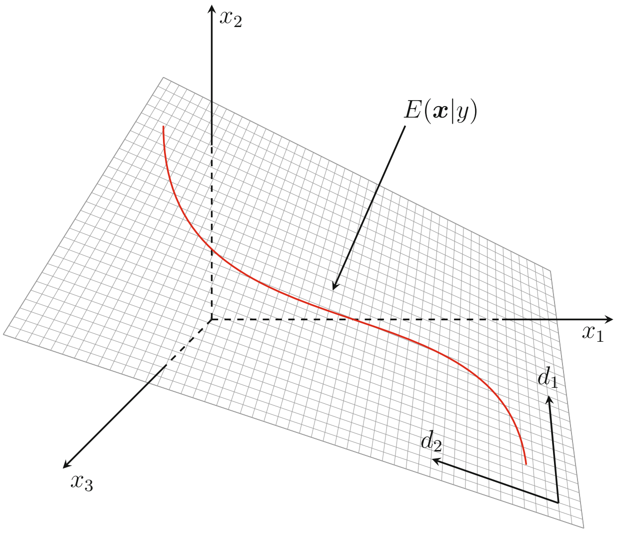


Figure 2 Inverse regression curve in a three-dimensional space

Similar unbiasedness property can be proved for our cluster-based SIR.

Theorem 2.3 *Under certain linearity conditions,* $E(\boldsymbol{x}|y)$ *is contained in the space spanned by* $\Sigma_{\boldsymbol{x}}\Lambda\Gamma$*.*

Theorem 2.3 describes the desirable property that there is no estimation bias. The e.d.r.-space estimated by our CRSIR method contains the true inverse regression curve. The details of the proof are provided in the Appendix.

2.3 Orthogonalization

For a given dataset $\boldsymbol{X}$ with dimension $N\times T$, and clusters $\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_c$, the correlations between these clusters need to be reduced to zero, to achieve cluster-wise independence. QR decomposition along with projection operators is used to perform the orthogonalization.

To begin with, use QR decomposition to find the orthogonal bases of the first cluster $\boldsymbol{X}_1$, named as $\boldsymbol{Q}_1$. Next, project the second cluster $\boldsymbol{X}_2$ onto the space of span$\{\boldsymbol{Q}_1\}^{\perp}$, which is the orthogonal complement of the space spanned by $\boldsymbol{X}_1$, named as $\boldsymbol{X}_2^*$,

$$\boldsymbol{X}_2^*=(\boldsymbol{I}-\boldsymbol{Q}_1\boldsymbol{Q}_1^{\mathrm{T}})\boldsymbol{X}_2. \tag{3}$$

Then use QR decomposition again to find the orthogonal bases of $\boldsymbol{X}_2^*$, named as $\boldsymbol{Q}_2$, and project $\boldsymbol{X}_3$ onto the space of span$\{\boldsymbol{Q}_1, \boldsymbol{Q}_2\}^{\perp}$, named as $\boldsymbol{X}_3^*$. Keep doing such process till the last cluster $\boldsymbol{X}_c$, we will get a new sequence of clusters $\boldsymbol{X}_1, \boldsymbol{X}_2^*, \cdots, \boldsymbol{X}_c^*$, in which every two clusters are orthogonal, and the new sequence contains all the information of the original dataset $\boldsymbol{X}$.

2.4 Regularization

Due to the high correlations between the series within each cluster, the covariance matrices of each cluster $\Sigma_{\boldsymbol{x}_i}$ are ill-conditioned, which make them hard to be inversed. We suggest a regularized version of the covariance matrix to overcome this issue (Friedman[13]).

$$\Sigma_{\boldsymbol{x}_i}(\tau) = (1-\tau)\Sigma_{\boldsymbol{x}_i} + \tau \frac{\mathrm{tr}\Sigma_{\boldsymbol{x}_i}}{N_i} I_{N_i}, \tag{4}$$

where $\tau \in [0, 1]$ is the shrinkage parameter. This is similar to the ridge version proposed by Zhong, et al.[6], which replaces $\Sigma_{\boldsymbol{x}_i}$ with $\Sigma_{\boldsymbol{x}_i} + \tau I_{N_i}$.

The shrinkage parameter τ can be chosen by cross-validation. Note when $\tau = 1$, the regularized covariance matrix will degenerate to a diagonal matrix whose diagonal elements are the means of the eigenvalues of $\Sigma_{\boldsymbol{x}_i}$. In such case, the chosen e.d.r.-direction is one of the input series, and the other series, which may also contain information for the predictors, are discarded.

2.5 Comparison Between CRSIR and SIR

Before applying the proposed CRSIR method to real data, consider the following simulated example first, for comparing the performance of CRSIR and SIR methods.

We choose γ clusters of predictors with cluster size 10, say, $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_\gamma$, which are independent and identically distributed (i.i.d.) with multivariate normal distribution $N(0, \Sigma)$, where Σ is a 10×10 covariance matrix with 1 at diagonal and 0.9 at off-diagonal.

The response y is simulated using the following formula

$$y = \sum_{j=1}^{\gamma} j \times \boldsymbol{x}_j + \boldsymbol{e},$$

where the random error $\boldsymbol{e}$ is independent to $\boldsymbol{x}_i$'s, and follows normal distribution $N(0, 0.1)$.

For simplicity, as well as keeping consistent with our following example, root mean square error (RMSE) is considered as a criterion to evaluate both in-sample prediction and out-of-sample forecasting.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{T} \left(\widehat{y}_i - y_i\right)^2 \Big/ T}, \tag{5}$$

where $\widehat{y}_i$ is the ith predicted value of the response, y_i is the ith observed value, and T is the number of observations.

We simulate 600 observations, in which 300 of them are used as training data and the others are used as testing data, at each run under above conditions. In CRSIR, the parameters c and

τ are chosen to minimize the in-sample RMSE for each run. Table 2.5 presents the means and standard deviations (in the parentheses) for the RMSE of SIR and CRSIR across 100 runs for several cluster numbers, and the median of the corresponding optimal c and τ.

Table 1 Simulation results for CRSIR and SIR

	SIR		CRSIR			
	In-sample	Out-of-sample	In-sample	Out-of-sample	median(c)	median(τ)
$\gamma = 1$	1.64(0.40)	1.65(0.41)	1.63(0.29)	1.65(0.35)	2	0.52
$\gamma = 2$	3.25(0.63)	3.41(0.69)	3.09(0.42)	3.22(0.46)	2	0.56
$\gamma = 5$	8.24(1.06)	9.07(1.34)	6.34(0.46)	6.56(0.45)	5	0.71
$\gamma = 10$	16.62(2.01)	20.41(2.61)	10.08(0.76)	10.55(0.94)	10	0.74

From Table 1, it is clear that CRSIR has better results than SIR for both in-sample prediction and out-of-sample forecasting. The CRSIR has similar RMSEs with smaller standard deviations when the number of clusters degenerates to 1, but it appears to be superior when the number of clusters increases. Besides, CRSIR outstands itself in out-of-sample forecasting, RMSEs for the testing data are almost the same as the one for training data, while SIR has much larger out-of-sample RMSEs. In addition, our other simulations, which are not presented here, show that CRSIR performs even better when the sample size T decreases to N.

3 Empirical Application

3.1 Dataset and Method

The dataset we use is Stock and Watson[3] dataset, which contains 143 quarterly macroeconomic variables from 13 economic categories, such as gross domestic product (GDP), industrial production (IP), employment, price indexes, interest rates, etc. We use 109 of them as explanatory variables, since the other 34 are just high-level aggregates of the 109. All 143 variables are used for forecasting purposes.

Following Stock and Watson's data transformation methodology, first differences of logarithms, first differences, and second differences of logarithms are used for real quantity variables, nominal interest rate variables, and price series, respectively. The correlation plot of the 109 predictor series after logarithm and/or differences is showed in Figure 3, which demonstrates that there do exist some highly correlated blocks. Therefore, our cluster-based method is necessary for this dataset.

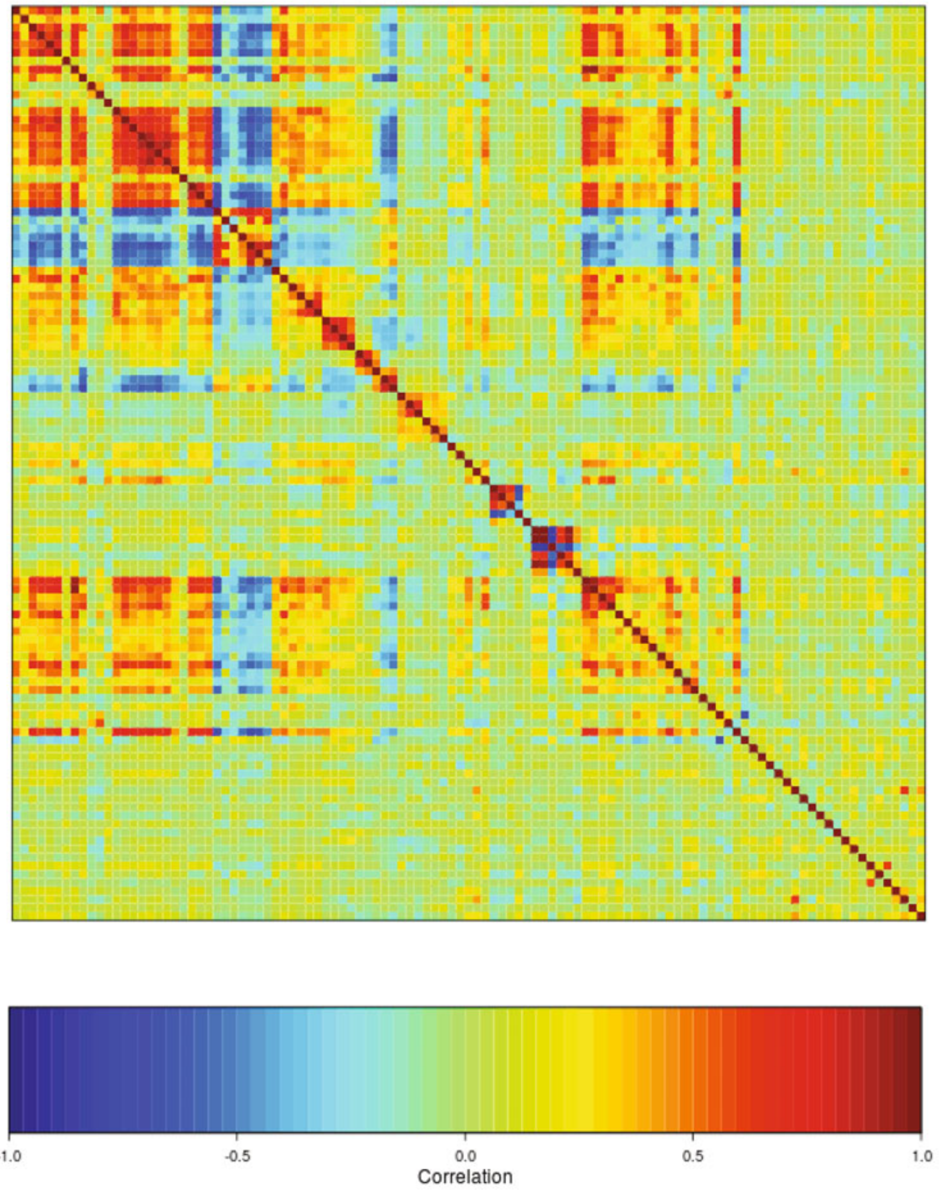


Figure 3 Plot of correlations of the 109 predictor series

For the purpose of comparison, similar rolling pseudo out-of-sample forecasting simulation as in [3] is used, as well as the cross validation for choosing c and τ. In general, starting from 1985 to 2008, at each date t, using the data prior to t to predict the forecasted variable y at h date ahead, which is denoted as $\widehat{y}_{t+h}$. The main steps can be described as follows:

1) Use the formula given by Stock and Watson[3]. Table 2 to transform all the series and screen for outliers.

2) At each date t, use cross-validation, which is described below, to the most recent 100

observations to choose the parameter c and τ in CRSIR based on mean square error.

3) Use the chosen $\widehat{c}$ and $\widehat{\tau}$ with the data prior to t to predict $\widehat{y}_{t+h}$ by CRSIR.

4) Calculate the RMSE for the forecasting procedure,

$$\text{RMSE} = \sqrt{\sum_{t=1}^{T} \left(y_{t+h} - \widehat{y}_{t+h}\right)^2 \Big/ T}.$$

The steps for cross-validation are described as follows:

(i) Regress y_{t+h} and x_t on the autoregressive terms $1, y_t, y_{t-1}, y_{t-2}, y_{t-3}$, in order to eliminate the autoregressive effect. Denote the residuals as $\widetilde{y}_{t+h}$ and $\widetilde{x}_t$.

(ii) Let $\Im(t) = \{1, 2, \cdots, t-2h-3, t+2h+3, \cdots, 100\}$, at each date $t = 1, 2, \cdots, 100-h$, find the e.d.r-directions and linear regression model using CRSIR and observations $\widetilde{y}_i$ and $\widetilde{x}_i,\ i \in \Im(t)$.

(iii) Use the e.d.r-directions and linear regression model from the above step at date t to predict $\widetilde{y}_{t+h}$.

(iv) For fixed h, parameters (c, τ) are chosen by minimizing the sum of squared forecasting error,

$$(\widehat{c}, \widehat{\tau}) = \text{argmin} \frac{1}{100-h} \sum_{t=1}^{100-h} \left(\widetilde{y}_{t+h} - \widehat{\widetilde{y}}_{t+h}\right)^2.$$

3.2 Results

We compare our method with the dynamic factor model using the first five principle components (DFM-5), which was claimed to be no worse than any other shrinkage methods according to Stock and Watson[3]. Besides, autoregressive model of order 4 (AR(4)) is used as a benchmark, and all RMSEs are recorded as the ratio relative to AR(4), smaller relative RMSE indicates better forecasting performance.

Table 2 presents the number of series with smaller RMSEs than AR(4) model for CRSIR and DFM-5. We can see that for forecasting period $h = 1$, if CRSIR is used, there are 97 series out of 143 have smaller RMSEs than the benchmark AR(4) model. If DFM-5 is used, only 85 series out of 143 have smaller RMSEs than AR(4) model. The differences become even larger for big forecasting period, when $h = 4$ the number of series of CRSIR increases to 115 while the number of DFM-5 decreases to 53.

Table 2 Number of series with smaller RMSE than AR(4) model

	DFM-5	**CRSIR**
$h = 1$	85	97
$h = 2$	59	109
$h = 4$	53	115

Table 3 presents the distributions of the RMSEs for AR(4), DFM-5, and CRSIR methods. When $h = 1$, the first quartile of the relative RMSE of CRSIR is just 0.768, which is much smaller than the relative RMSE of DFM-5 (0.961), and the median relative RMSE of CRSIR is 0.907, while DFM-5 has 0.993. When $h = 2$ and $h = 4$, CRSIR improves the forecasting results of AR(4) for more than 3/4 of the series. The relative RMSEs of CRSIR at first, second, and third quartile are all smaller than those of DFM-5.

From Tables 2 and 3, one can tell that CRSIR improves the forecasting results significantly compared to the DFM-5 method, especially for longer forecasting period.

Table 3 Distributions of relative RMSEs by Pseudo out-of-sample forecasting

(a) $h = 1$

Method	**Percentiles**				
	0.050	0.250	0.500	0.750	0.950
AR(4)	1.000	1.000	1.000	1.000	1.000
DFM-5	0.874	0.961	0.993	1.022	1.089
CRSIR	0.621	0.768	0.907	1.048	1.372

(b) $h = 2$

Method	**Percentiles**				
	0.050	0.250	0.500	0.750	0.950
AR(4)	1.000	1.000	1.000	1.000	1.000
DFM-5	0.882	0.976	1.010	1.044	1.125
CRSIR	0.652	0.759	0.865	0.991	1.186

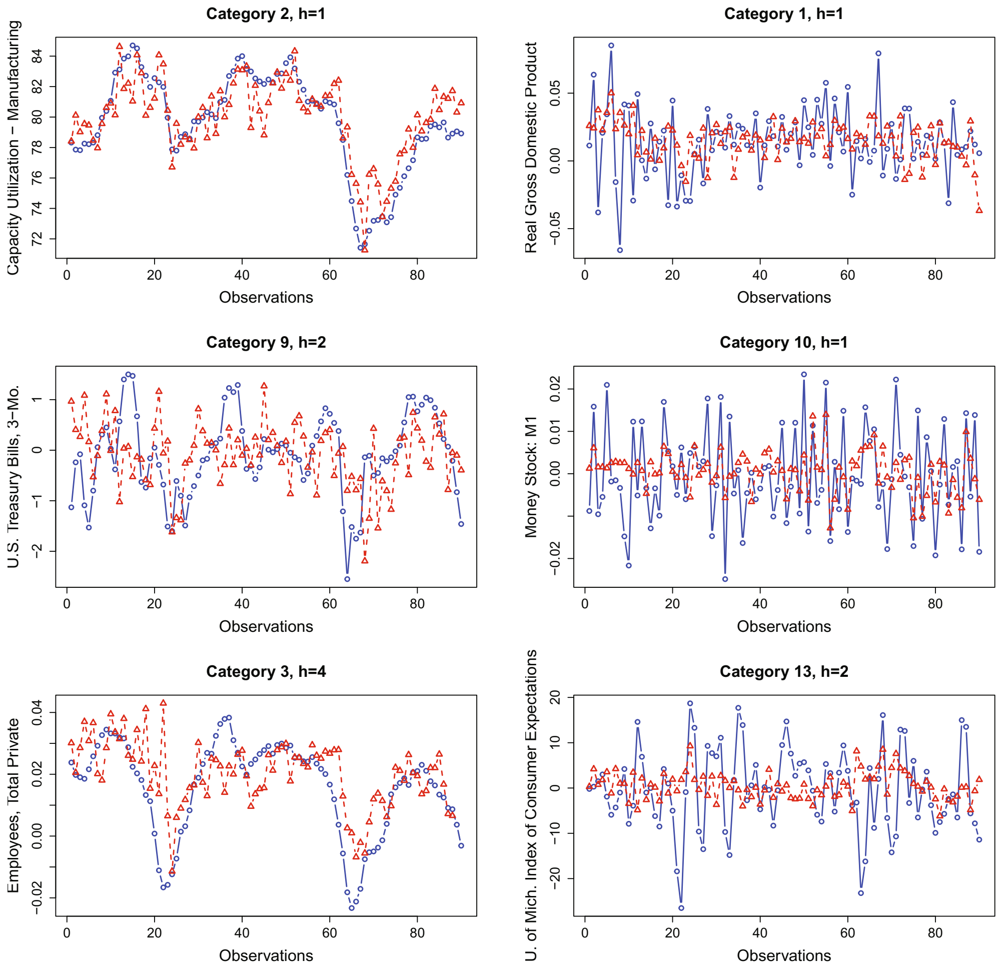
(c) $h = 4$

Method	**Percentiles**				
	0.050	0.250	0.500	0.750	0.950
AR(4)	1.000	1.000	1.000	1.000	1.000
DFM-5	0.903	0.980	1.020	1.058	1.138
CRSIR	0.648	0.736	0.827	0.940	1.220

Table 4 presents the median RMSEs relative to AR(4) model by category via cross-validation. Column "S&W" reports the smallest relative RMSE Stock and Watson got using DFM-5 and other shrinkage methods in their 2011 paper. Comparing all these results, CRSIR method has smaller median relative RMSEs for more than 70% of these categories among three forecasting period, which demonstrates its superiority again.

Table 4 Median relative RMSE for forecasting by category of series

	$h=1$			$h=2$			$h=4$		
Category	**DFM-5**	**S&W**	**CRSIR**	**DFM-5**	**S&W**	**CRSIR**	**DFM-5**	**S&W**	**CRSIR**
1. GDP Components	0.905	0.905	1.079	0.907	0.870	0.807	0.906	0.906	0.839
2. Industrial Production	0.882	0.882	0.669	0.861	0.852	0.694	0.827	0.827	0.745
3. Employment	0.861	0.861	0.849	0.861	0.859	0.803	0.844	0.842	0.823
4. Unempl. Rate	0.800	0.799	0.771	0.750	0.723	0.723	0.762	0.743	0.647
5. Housing	0.936	0.897	1.220	0.940	0.902	1.081	0.926	0.882	0.807
6. Inventories	0.900	0.886	0.856	0.867	0.867	0.764	0.856	0.856	0.784
7. Prices	0.980	0.970	0.865	0.977	0.961	0.892	0.963	0.948	0.797
8. Wages	0.993	0.938	0.967	0.999	0.919	0.960	1.019	0.931	1.031
9. Interest Rates	0.980	0.946	0.849	0.952	0.928	0.892	0.956	0.949	0.822
10. Money	0.953	0.926	1.000	0.933	0.921	0.950	0.909	0.909	0.927
11. Exchange Rates	1.015	0.981	0.974	1.015	0.980	1.108	1.036	0.965	1.150
12. Stock Prices	0.983	0.983	0.840	0.977	0.955	0.893	0.974	0.961	1.039
13. Cons. Exp.	0.977	0.977	0.765	0.963	0.960	1.082	0.966	0.955	0.963



(a) From CRSIR Favored Categories (b) From CRSIR No-Favored Categories

Figure 4 Plots of the forecasting values (△) vs. real observations (○) from 1985 to 2008

Table 4 also indicates the performance of CRSIR varied across categories. It has outstanding performance for some categories, such as Industrial Production, Unemployment Rate, Inventories, Interest Rates, etc. But it does not work well for some others, such as Housing, Money, Exchange Rates. Figure 5 plots six series from both CRSIR favored and no-favored categories. Three of them in Figure 4(a) are from CRSIR favored categories and three of them in Figure 4(b) are from CRSIR no-favored categories. From these plots, one can see that the responses of the CRSIR no-favored series are quite disordered. They are more like white noises, the variations are big but the changes of $\boldsymbol{x}$ means are not distinct. The inverse regression method is aimed to detect the variation of $E(\boldsymbol{x}|y)$. If the conditional expectations of $\boldsymbol{x}$ do not have much difference for different values of y, the estimation for the e.d.r.-directions will be inaccurate, and will lead to the poor performance on forecasting.

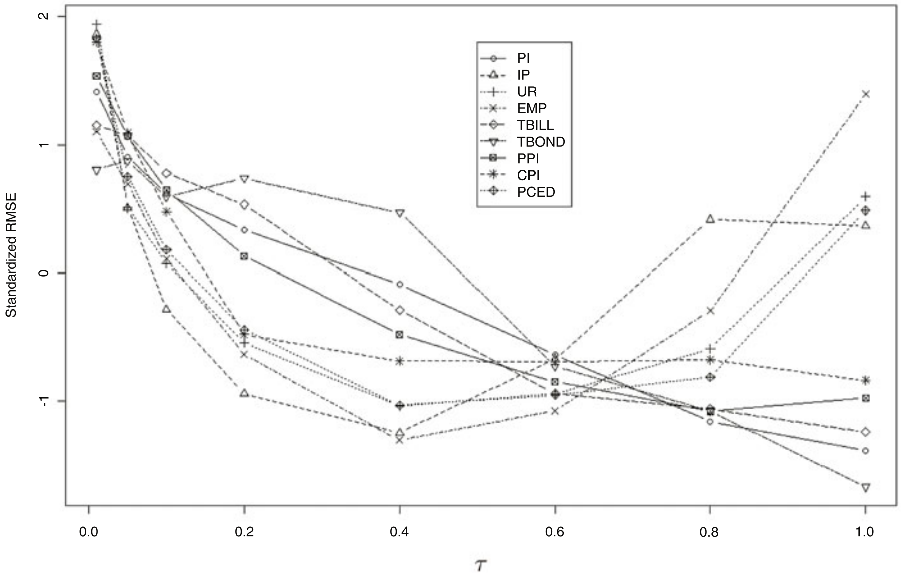


Figure 5 RMSE vs. τ

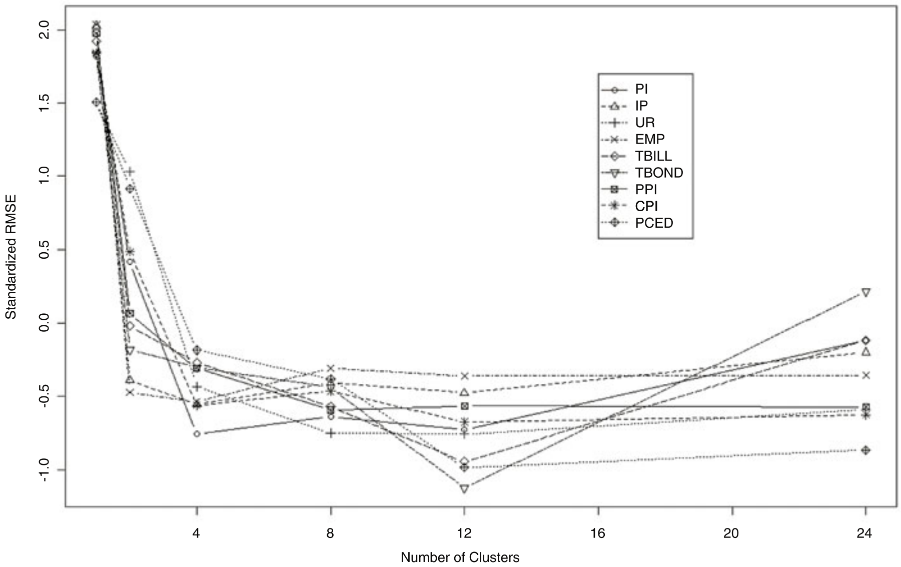


Figure 6 RMSE vs. number of clusters

Six series are reported as illustrations to show how the estimated RMSE changes when τ or c changes. They are real average hourly earnings (PI), industrial production index (IP), unemployment rate (UR), employees (EMP), 3 months treasury bills (TBILL), and 10 years treasury const maturities (TBOND). Figure 6 shows the plot of their RMSEs with the values of shrinkage parameter τ for $h = 2$ and $c = 10$, Figure 6 shows the plot of their RMSEs with the number of clusters c for $h = 2$ and $\tau = 0.5$. RMSEs in both figures are standardizes to the same scale for comparing purposes. These two figures confirm that the clustering and regularized approach do enhance the performance for the regular SIR method, and for this dataset, optimal τ is between 0.4 to 0.8, and optimal c is between 8 to 12.

Figure 7 presents the pair plots of the forecasting series y with the first three e.d.r.-directions estimated by CRSIR for one of the series. The other series in all horizons had similar results. It shows that the relation between the forecasting series and e.d.r.-directions is close to linear, along with the independence among e.d.r.-directions, it is reasonable to use linear model with OLS to predict the values of y.

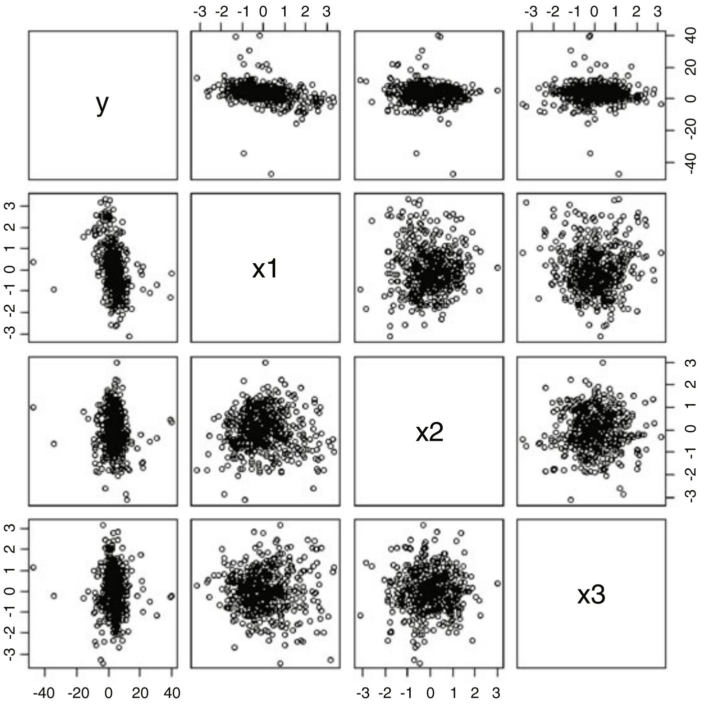


Figure 7 y vs. the first three e.d.r.-directions x_1, x_2, and x_3

4 Conclusion and Discussion

Sliced inverse regression now becomes a popular dimension reduction method in computer science, engineering and biology. In this article, we bring it to macroeconomic forecasting model when there are a large number of predictors and high collinearity. Compared to the classical dynamic factor model, SIR retrieves information not only from the predictors but also from the response. Moreover, our cluster-based regularized SIR has the ability to handle highly collinearity or "$T < N$" data. The simulation confirms that it offers a lot of improvements over DFM-5 model on the macroeconomic data set.

After finding the CRSIR variates, we use linear models for forecasting the responses y, because scatter plots for CRSIR variates and y values show strong linear relationships, and the results are desirable. But one may use polynomials, splines, Lasso, or some other more advanced regression techniques for different cases to get better fitting results.

Based on its basic idea, there are more than one generalizations of SIR using higher order inverse moments. For instance, SAVE[14], SIR-II[15], DR[16], and SIMR[17]. Our cluster-based algorithm can also be applied to these methods for highly collinearity data, and good performance is expected.

Above all, we can conclude that the cluster-based regularized sliced inverse regression is a powerful tool in forecasting using many predictors. It may not be limited in macroeconomic forecasting, and can also be applied to dimension reduction or variable selection problems in social science, microarray analysis, or clinical trails when the dataset is large and highly correlated.

References

[1] Boivin J and Ng S, Understanding and comparing factor-based forecasts, *International Journal of Central Banking*, 2005.

[2] Eickmeier S and Ziegler C, How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach, *Journal of Forecasting*, 2008, **27**(3): 237–265.

[3] Stock J H and Watson M W, Generalized shrinkage methods for forecasting using many predictors, *Journal of Business & Economic Statistics*, 2012, **30**(4): 481–493.

[4] Duan N and Li K C, Slicing regression: A link-free regression method, *The Annals of Statistics*, 1991, **19**(2): 505–530.

[5] Li K C, Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, 1991, **86**(414): 316–327.

[6] Zhong W, Zeng P, Ma P, Liu J S, and Zhu Y, Rsir: Regularized sliced inverse regression for motif discovery, *Bioinformatics*, 2005, **21**(22): 4169–4175.

[7] Li L and Yin X, Sliced inverse regression with regularizations, *Biometrics*, 2008, **64**(1): 124–131.

[8] Li L, Cook R D, and Tsai C L, Partial inverse regression, *Biometrika*, 2007, **94**(3): 615–625.

[9] Li K C, High dimensional data analysis via the sir/phd approach, 2000.

[10] Ward J H, Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, 1963, **58**(301): 236–244.

[11] Eaton M L, A characterization of spherical distributions, *Journal of Multivariate Analysis*, 1986, **20**(2): 272–276.

[12] Hall P and Li K C, On almost linearity of low dimensional projections from high dimensional data, *The Annals of Statistics*, 1993, **21**(2): 867–889.

[13] Friedman J H, Regularized discriminant analysis, *Journal of the American Statistical Association*, 1989, **84**(405): 165–175.

[14] Cook R D and Weisberg S, Discussion of Li (1991), *Journal of the American Statistical Association*, 1991, **86**: 328–332.

[15] Li K C, Sliced inverse regression for dimension reduction: Rejoinder, *Journal of the American Statistical Association*, 1991, **86**(414): 337–342.

[16] Li B and Wang S, On directional regression for dimension reduction, *Journal of the American Statistical Association*, 2007, **102**(479): 997–1008.

[17] Ye Z and Yang J, Sliced inverse moment regression using weighted chi-squared tests for dimension reduction, *Journal of Statistical Planning and Inference*, 2010, **140** (11): 3121–3131.

Appendix

Assume the following linearity conditions.

Linearity Condition A.1 For any $\boldsymbol{b} \in \mathbb{R}^{N_i}$, the conditional expectation $E(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_i|\boldsymbol{\theta}_j^{(i)T}\boldsymbol{x}_i)$ is linear in $\boldsymbol{\theta}_j^{(i)T}\boldsymbol{x}_i$, $j = 1, 2, \cdots, k_i$.

Linearity Condition A.2 For any $\boldsymbol{b} \in \mathbb{R}^N$, the conditional expectation $E(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}|\Lambda^{\mathrm{T}}\boldsymbol{x})$ is linear in $\boldsymbol{\lambda}_1^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{\lambda}_2^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{\lambda}_m^{\mathrm{T}}\boldsymbol{x}$.

Linearity Condition A.3 For any $\boldsymbol{b} \in \mathbb{R}^m$, the conditional expectation $E(\boldsymbol{b}^{\mathrm{T}}\Lambda^{\mathrm{T}}\boldsymbol{x}|\Gamma^{\mathrm{T}}\Lambda^{\mathrm{T}}\boldsymbol{x})$ is linear in $\boldsymbol{\gamma}_1^{\mathrm{T}}\Lambda^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{\gamma}_2^{\mathrm{T}}\Lambda^{\mathrm{T}}\boldsymbol{x}, \cdots, \boldsymbol{\gamma}_v^{\mathrm{T}}\Lambda^{\mathrm{T}}\boldsymbol{x}$.

Conditions A.1 and A.2 are satisfied when all the $\boldsymbol{x}$'s have elliptical symmetric distribution, especially the multivariate normal distribution (Eaton[11]). Condition A.3 is also satisfied when all the $\Lambda^{\mathrm{T}}\boldsymbol{x}$ have elliptical symmetric distribution, which is true because all the elliptical symmetric distributed $\boldsymbol{x}$'s have been standardized to the same scale.

Li's Theorem 2.2 can be restated as following for each cluster when $E(\boldsymbol{x}) = 0$.

Theorem A.4 ([5]) *Under Linearity Condition* A.1*,* $E(\boldsymbol{x}_i|y)$ *is contained in the space spanned by* $\Sigma_{\boldsymbol{x}_i}\boldsymbol{\theta}_j^{(i)}$, $j = 1, 2, \cdots, k_i$.

Furthermore, it's not hard to see that,

Corollary A.5 *Under Linearity Condition* A.2*,* $E(\boldsymbol{x}|y)$ *is contained in the space spanned by* $\Sigma_{\boldsymbol{x}}\boldsymbol{\lambda}_1, \Sigma_{\boldsymbol{x}}\boldsymbol{\lambda}_2, \cdots, \Sigma_{\boldsymbol{x}}\boldsymbol{\lambda}_m$.

Corollary A.6 *Under Linearity Condition* A.3*,* $E(\Lambda^{\mathrm{T}}\boldsymbol{x}|y)$ *is contained in the space spanned by* $\Sigma_{\Lambda^{\mathrm{T}}\boldsymbol{x}}\boldsymbol{\gamma}_1, \Sigma_{\Lambda^{\mathrm{T}}\boldsymbol{x}}\boldsymbol{\gamma}_2, \cdots, \Sigma_{\Lambda^{\mathrm{T}}\boldsymbol{x}}\boldsymbol{\gamma}_v$.

Based on the above results, we can conclude that

Theorem A.7 *Under Linearity Conditions* A.1, A.2*, and* A.3*,* $E(\boldsymbol{x}|y)$ *is contained in the space spanned by* $\Sigma_{\boldsymbol{x}}\Lambda\Gamma$.

Proof Li[9] proved Theorem 2.2, which is the same as Corollary A.5 in different notations, by showing that $E(\boldsymbol{x}|y)$ can be written as

$$E(\boldsymbol{x}|y) = \Sigma_{\boldsymbol{x}} \Lambda \kappa_1(y),$$

where $\kappa_1(y) = (\Lambda^{\mathrm{T}} \Sigma_{\boldsymbol{x}} \Lambda)^{-1} E(\Lambda^{\mathrm{T}} \boldsymbol{x}|y)$.

Similarly, under Condition A.3,

$$E(\Lambda^{\mathrm{T}} \boldsymbol{x}|y) = \Sigma_{\Lambda^{\mathrm{T}} \boldsymbol{x}} \Gamma \kappa_2(y),$$

where $\kappa_2(y) = (\Gamma^{\mathrm{T}} \Sigma_{\Lambda^{\mathrm{T}} \boldsymbol{x}} \Gamma)^{-1} E(\Gamma^{\mathrm{T}} \Lambda^{\mathrm{T}} \boldsymbol{x}|y)$ and $\Sigma_{\Lambda^{\mathrm{T}} \boldsymbol{x}} = \Lambda^{\mathrm{T}} \Sigma_{\boldsymbol{x}} \Lambda$.

Therefore,

$$\begin{aligned} E(\boldsymbol{x}|y) = \Sigma_{\boldsymbol{x}} \Lambda \kappa_1(y) &= \Sigma_{\boldsymbol{x}} \Lambda (\Lambda^{\mathrm{T}} \Sigma_{\boldsymbol{x}} \Lambda)^{-1} E(\Lambda^{\mathrm{T}} \boldsymbol{x}|y) \\ &= \Sigma_{\boldsymbol{x}} \Lambda (\Lambda^{\mathrm{T}} \Sigma_{\boldsymbol{x}} \Lambda)^{-1} \Lambda^{\mathrm{T}} \Sigma_{\boldsymbol{x}} \Lambda \Gamma \kappa_2(y) \\ &= \Sigma_{\boldsymbol{x}} \Lambda \Gamma \kappa_2(y). \end{aligned}$$

That implies that $E(\boldsymbol{x}|y)$ is in the e.d.r. space spanned by $\Sigma_{\boldsymbol{x}} \Lambda \Gamma$.