Dynamic Cluster-based Regularized Sliced Inverse Regression

for Forecasting Microeconomics Variables

Yue Yu*

with Prof. Jie Yang* and Prof. Zhihong Chen†

∗: University of Illinois at Chicago
†: University of International Business and Economics, China

Oct. 6, 2010

Contents

UIC University of Illinois at Chicago

Sliced Inverse Regression (SIR)

- Ker-Chau Li (1991) introduced the following model:

$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon) \qquad (1)$$

Sliced Inverse Regression (SIR)

- Ker-Chau Li (1991) introduced the following model:

$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon) \tag{1}$$

- y is an univariate output variable;
- The dimension of $\mathbf{x}$ is p;
- The random error ϵ is independent of $\mathbf{x}$;
- The space $\mathfrak{B}$ generated by $\beta_1, \ldots, \beta_K$ is called the **effective dimension reduction** (e.d.r.) space.

UIC University of Illinois at Chicago

Sliced Inverse Regression (SIR)

- Ker-Chau Li (1991) introduced the following model:

$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon) \tag{1}$$

- Model (1) is equivalent to:
  - The conditional distribution of y given $\mathbf{x}$ depends on $\mathbf{x}$ only through the K dimensional variable $(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x})$;
  - Conditional on $(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x})$, y and $\mathbf{x}$ are independent.

Sliced Inverse Regression (SIR)

- Unlike other common methods;

Sliced Inverse Regression (SIR)

- Unlike other common methods;

Figure: General Regression Model

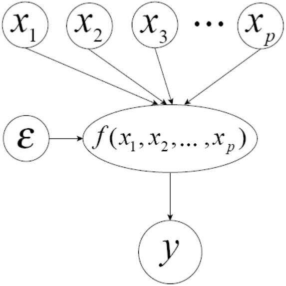
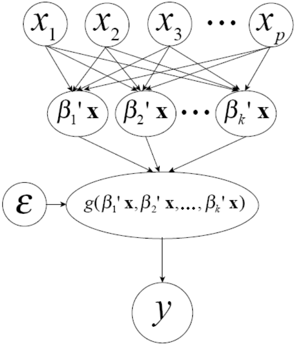


Figure: Effective Dimension Reduction



Sliced Inverse Regression (SIR)

- Unlike other common methods;
- SIR reverses the role of $\mathbf{x}$ and y.
  - Instead of estimating the forward regression function

$$\eta(\mathbf{x}) = E(y|\mathbf{x}),$$

an inverse regression function is considered:

$$\xi(\mathbf{y}) = E(\mathbf{x}|y).$$

UIC University of Illinois at Chicago

Algorithm of SIR

- Standardize predictors x_i:

$$z_i = \hat{\Sigma}_x^{-1/2}(x_i - \bar{\mathbf{x}}), \text{ where } \hat{\Sigma}_x = \sum_{i=1}^{p}(x_i - \bar{\mathbf{x}})(x_i - \bar{\mathbf{x}})'/p;$$

- Sort the values of y and then partition them into H slices;

- Distribute z_i into H slices and compute their covariance for slice means:

$$\Sigma_\xi = \sum_{h=1}^{H} \hat{\rho}_h \bar{z}_h \bar{z}_h',$$

  where $\hat{\rho}_h$ is the proportion of observations falling into slice h, and $\bar{z}_h = \sum_{i=1}^{p} I_{z_i \in h} z_i / n_i$;

- Find the eigenvector of Σ_ξ, $\hat{\eta}_1, \hat{\eta}_2, \cdots, \hat{\eta}_K$, the e.d.r. directions are

$$\hat{\beta}_k = \hat{\Sigma}_x^{-1/2}\hat{\eta}_k, \quad k = 1, 2, \cdots, K.$$

Fisher Consistency of SIR

Linearity Condition

For any $b \in \mathbb{R}^p$, the conditional expectation $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x})$ is linear in $\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}$.

Fisher Consistency of SIR

Linearity Condition

For any $b \in \mathbb{R}^p$, the conditional expectation $E(b'\mathbf{x}|\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x})$ is linear in $\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x}$.

- Eaton (1986) showed when X is elliptically symmetrically distributed, and particularly, when X follows a multivariate normal distribution, the linearity condition holds.
- Hall and Li (1993) showed that this is not a restrictive assumption, because it holds to a reasonable approximation as p increases.

Fisher Consistency of SIR

Theorem by Ker-Chau Li (1991)

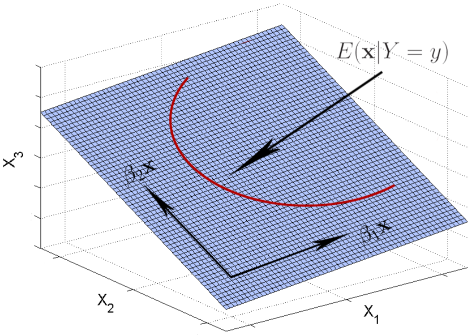
Assume Linearity Condition, the standardized **inverse regression curve** $\mathbf{E}(\mathbf{x}|\mathbf{Y} = \mathbf{y})$ is contained in the space spanned by the e.d.r. directions $\beta_i,\ i = 1, \ldots, K$.

University of Illinois at Chicago

Fisher Consistency of SIR

Theorem by Ker-Chau Li (1991)

Assume Linearity Condition, the standardized **inverse regression curve $\mathbf{E}(\mathbf{x}|\mathbf{Y}=\mathbf{y})$** is contained in the space spanned by the e.d.r. directions $\beta_i,\ i=1,\ldots,K$.



Test of Eigenvalues (ν)

Theorem

If $\mathbf{x}$ is normally distributed, then $n(p-K)\bar{\nu}_{p-K}$ follows a χ^2 distribution with $(p-K)(H-K-1)$ degrees of freedom asymptotically.

University of Illinois at Chicago

Test of Eigenvalues (ν)

Theorem

If $\mathbf{x}$ is normally distributed, then $n(p-K)\bar{\nu}_{p-K}$ follows a χ^2 distribution with $(p-K)(H-K-1)$ degrees of freedom asymptotically.

- We can decide how many directions are used by using a sequential p–value:

$$p\text{–value} = P\left\{\chi^2_{(p-j)(H-j-1)} \geqslant n(p-j)\bar{\nu}_{p-j}\right\}.$$

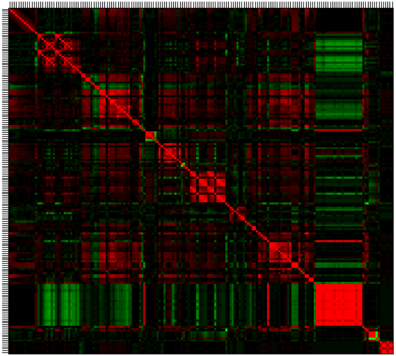
- Begin with $j = 0$, if p–value is less than say 0.05, then there are at least j+1 directions.

Microeconomics Data

- We use the data from the study of Zhihong Chen, et al. (2010). The data are monthly collected from 1964–01 to 2007–12 and contains 207 predictors totally, which include 15 main categories like real prices indexes; employment and hours; manufacturing and trade sales; etc.

Microeconomics Data

- We use the data from the study of Zhihong Chen, et al. (2010). The data are monthly collected from 1964–01 to 2007–12 and contains 207 predictors totally, which include 15 main categories like real prices indexes; employment and hours; manufacturing and trade sales; etc.



Normalization

- Relative difference: $(x(t) - x(t-1))/x(t-1)$;
- Box–Cox power transformation;
- Box–Cox power transformation after relative difference.

Microeconomics Data

Normalization

- Relative difference: $(x(t) - x(t-1))/x(t-1)$;
- Box–Cox power transformation;
- Box–Cox power transformation after relative difference.

- Not all the variables are available from 1964–01, remove the variables have too many NAs.
- Standardize the data after normalization.

Autoregressive Model

- We have to consider the possibly time-lagged relations between the variables.

Autoregressive Model

- We have to consider the possibly time-lagged relations between the variables.
- Assume a multivariate nonlinear autoregressive model.

$$y(t) = h(\mathbf{x}(t), y(t-1), y(t-2), \ldots, y(t-l), \varepsilon)$$

Autoregressive Model

- We have to consider the possibly time-lagged relations between the variables.

- Assume a multivariate nonlinear autoregressive model.

$$y(t) = h(\mathbf{x}(t), y(t-1), y(t-2), \ldots, y(t-l), \varepsilon)$$

- Checking:
  - Lag plot for the residuals;
  - Durbin–Watson h test / Breusch–Godfrey Lagrange multiplier test.

Multicollinearity Problem

The e.d.r. directions are

$$\hat{\beta}_k = \hat{\Sigma}_x^{-1/2} \hat{\eta}_k, \quad k = 1, 2, \cdots, K.$$

- Multicollinearity of our variables make the covariance matrix ill-conditioned.
- The inverse of the covariance matrix or eigenvalues/vectors computation is sensitive and have potential accuracy problem.
- Cause the false and unstable selection of the e.d.r. directions.

Cluster-based Regularized SIR

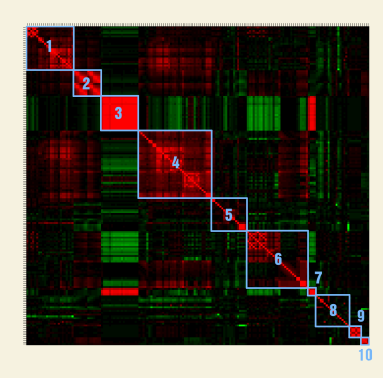
Algorithm of Cluster-based Regularized SIR

- Partition $\mathbf{x}$ into c clusters based on the correlation matrix (clustering method: complete linkage/farthest neighbor);

Cluster-based Regularized SIR

Algorithm of Cluster-based Regularized SIR

- Partition $\mathbf{x}$ into c clusters based on the correlation matrix (clustering method: complete linkage/farthest neighbor);



Cluster-based Regularized SIR

Algorithm of Cluster-based Regularized SIR

- Partition $\mathbf{x}$ into c clusters based on the correlation matrix (clustering method: complete linkage/farthest neighbor);
- Perform **SIR/regularized SIR** in each cluster, choose the number of e.d.r. directions based on the χ^2 test;
- Combine all the e.d.r. directions chosen from the clusters;
- Perform another SIR to the pooled directions, choose the number of e.d.r. directions based on the χ^2 test;
- Predict y based on the e.d.r. directions chosen from the above step.
  - ◇ Any parametrical/nonparametrical model can be used.
  - ◇ We use linear regression.

Regularized SIR

- A shrunken version of $\hat{\Sigma}_x$ is used to overcome ill-condition of the covariance matrix. (Friedman, 1989)

$$\hat{\Sigma}_x(\tau) = (1-\tau)\hat{\Sigma}_x + \tau \frac{\mathrm{tr}\hat{\Sigma}_x}{p} I_p.$$

Regularized SIR

- A shrunken version of $\hat{\Sigma}_x$ is used to overcome ill-condition of the covariance matrix. (Friedman, 1989)

$$\hat{\Sigma}_x(\tau) = (1 - \tau)\hat{\Sigma}_x + \tau \frac{\mathrm{tr}\hat{\Sigma}_x}{p} I_p.$$

- Scrucca (2006) introduced a regularized SIR with the weighted average of both SIR and SIR-II methods.
  - SIR-II: gains information from variation on class variances instead of means.
  - It has poor performance for our dataset.

Fisher Consistency of Cluster-based SIR

- Without loss of generality, assume $E(\mathbf{x}) = 0$ and $\operatorname{cov}(\mathbf{x}) = I$;

UIC University of Illinois at Chicago

Fisher Consistency of Cluster-based SIR

- Without loss of generality, assume $E(\mathbf{x}) = 0$ and $\text{cov}(\mathbf{x}) = I$;
- Partition $\mathbf{x}$ to $\begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_c} \end{pmatrix}$, $\mathbf{x_i} = p_i \times n$, $\sum_{i=1}^{c} p_i = p$;

Fisher Consistency of Cluster-based SIR

- Without loss of generality, assume $E(\mathbf{x}) = 0$ and $\mathrm{cov}(\mathbf{x}) = I$;
- Partition $\mathbf{x}$ to $\begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_c} \end{pmatrix}$, $\mathbf{x_i} = p_i \times n$, $\sum_{i=1}^{c} p_i = p$;
- Within each cluster
  - Model:

$$y = g(\theta_1^{(i)\prime}\mathbf{x_i}, \theta_2^{(i)\prime}\mathbf{x_i}, \ldots, \theta_{k_i}^{(i)\prime}\mathbf{x_i}, \epsilon), \quad \theta_j^{(i)} : p_i \times 1.$$

Theorem

$E(\mathbf{x_i}|y)$ is in the e.d.r. space spaced by $\theta_j^{(i)}$, $j = 1, \ldots, k_i$.

Fisher Consistency of Cluster-based SIR

- Write $\lambda_{(i,j)} = \begin{pmatrix} \mathbf{0_1} \\ \theta_j^{(i)} \\ \mathbf{0_2} \end{pmatrix}$, and $\Lambda = (\lambda_{(1,1)}, \lambda_{(1,2)}, \dots)$,

  where so $\theta_j^{(i)\prime}\mathbf{x_i}$ can be written as $\lambda'_{(i,j)}\mathbf{x}$.

Corollary

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by Λ.

University of Illinois at Chicago

Fisher Consistency of Cluster-based SIR

- Write $\lambda_{(i,j)} = \begin{pmatrix} \mathbf{0_1} \\ \theta_j^{(i)} \\ \mathbf{0_2} \end{pmatrix}$, and $\Lambda = (\lambda_{(1,1)}, \lambda_{(1,2)}, \dots)$,

  where so $\theta_j^{(i)\prime}\mathbf{x_i}$ can be written as $\lambda'_{(i,j)}\mathbf{x}$.

Corollary

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by Λ.

- Perform SIR on the data $\Lambda'\mathbf{x}$, we have:

Theorem

$E(\Lambda'\mathbf{x}|y)$ is in the e.d.r. space spanned by $B = (\beta_i)$.

Fisher Consistency of Cluster-based SIR

Corollary

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by Λ.

Theorem

$E(\Lambda'\mathbf{x}|y)$ is in the e.d.r. space spanned by $B = (\beta_i)$.

UIC University of Illinois at Chicago

Fisher Consistency of Cluster-based SIR

Corollary

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by Λ.

Theorem

$E(\Lambda'\mathbf{x}|y)$ is in the e.d.r. space spanned by $B = (\beta_i)$.

Proposition

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by ΛB.

Fisher Consistency of Cluster-based SIR

Proposition

$E(\mathbf{x}|y)$ is in the e.d.r. space spanned by ΛB.

Proof:

Ker-Chau Li proved, since $E(\mathbf{x}|y)$ is in the e.d.r. space spanned by Λ, under linearity condition,

$$E(\mathbf{x}|y) = \Sigma_x \Lambda \kappa_1(y) = \Lambda \kappa_1(y),$$

where $\kappa_1(y) = (\Lambda'\Lambda)^{-1} E(\Lambda'\mathbf{x}|y)$. Similarly,

$$E(\Lambda'\mathbf{x}|y) = \Sigma_{\Lambda'\mathbf{x}} B \kappa_2(y),$$

where $\kappa_2(y) = (B'\Sigma_{\Lambda'\mathbf{x}} B)^{-1} E(B'\Lambda'\mathbf{x}|y)$. Therefore,

$$\begin{aligned} E(\mathbf{x}|y) = \Sigma_x \Lambda \kappa_1(y) &= \Lambda(\Lambda'\Lambda)^{-1} \Sigma_{\Lambda'\mathbf{x}} B \kappa_2(y) \\ &= \Lambda(\Lambda'\Lambda)^{-1} (\Lambda'\Lambda) B \kappa_2(y) \\ &= \Lambda B \kappa_2(y) \end{aligned}$$

$\square$

UIC University of Illinois at Chicago

Model Fitting

- Normalize, Standardize, and clean NAs;

Model Fitting

- Normalize, Standardize, and clean NAs;
- Forecast the response series value h ($h = 6$) months later, ie., forecast $y(t+h)$ by using $(\mathbf{x}(t), y(t), \ldots, y(t-l))$;

Model Fitting

- Normalize, Standardize, and clean NAs;
- Forecast the response series value h ($h = 6$) months later, ie., forecast $y(t+h)$ by using $(\mathbf{x}(t), y(t), \ldots, y(t-l))$;
- To compare with 13 methods introduced in Stock and Watson's paper *An Empirical Comparison of Methods for Forecasting Using Many Predictors* (2005), choose the forecast series:

Series	Abbreviation	Y_{t+h}^h	Y_t
Real Personal Income	PI	$(1200/h)\ln(Z_{t+h}/Z_t)$	$\Delta\ln(Z_t)$
Industrial Production	IP	$(1200/h)\ln(Z_{t+h}/Z_t)$	$\Delta\ln(Z_t)$
Unemployment Rate	UR	$(Z_{t+h} - Z_t)$	ΔZ_t
Employment	EMP	$(1200/h)\ln(Z_{t+h}/Z_t)$	$\Delta\ln(Z_t)$
3-Mth Tbill Rate	TBILL	$(Z_{t+h} - Z_t)$	ΔZ_t
10-Yr TBond Rate	TBOND	$(Z_{t+h} - Z_t)$	ΔZ_t
Producer Price Index	PPI	$1200[(1/h)\ln(Z_{t+h}/Z_t) - \Delta\ln(Z_t)]$	$\Delta^2\ln(Z_t)$
Consumer Price Index	CPI	$1200[(1/h)\ln(Z_{t+h}/Z_t) - \Delta\ln(Z_t)]$	$\Delta^2\ln(Z_t)$
PCE Deflator	PCED	$1200[(1/h)\ln(Z_{t+h}/Z_t) - \Delta\ln(Z_t)]$	$\Delta^2\ln(Z_t)$

Model Fitting Criterion

- To be consistent to the forecast data Stock and Watson (2005) used, we choose to use the data from 1964–01 to 2007–12, and start prediction from 1978–01.

Model Fitting Criterion

- To be consistent to the forecast data Stock and Watson (2005) used, we choose to use the data from 1964–01 to 2007–12, and start prediction from 1978–01.

- Consider the root mean square error (RMSE) as a criterion for model fitting. For example, if we choose $h = 6$ and have M predicted values, the RMSE is:

$$\mathrm{RMSE}_6 = \sqrt{\frac{1}{M} \sum_{M} \Big[\hat{y}(t+6) - y(t+6)\Big]^2}.$$

Choosing of Parameters

- Number of slices H;

- Number of lags l;

- Number of clusters c;
- Shrinkage parameter for regularization τ.

Choosing of Parameters

- Number of slices H;
  - It's not crucial, there are theoretical results (Li, 2000) indicating the SIR outputs do not change much for a wide range of H.
  - Choose $H = 10$.
- Number of lags l;
- Number of clusters c;
- Shrinkage parameter for regularization τ.

Choosing of Parameters

- Number of slices H;
  - It's not crucial, there are theoretical results (Li, 2000) indicating the SIR outputs do not change much for a wide range of H.
  - Choose $H = 10$.
- Number of lags l;
  - It's not crucial, usually all the lags will be selected into one cluster.
  - To be consistent to Stock and Watson (2005) paper, choose $l = 4$.
- Number of clusters c;
- Shrinkage parameter for regularization τ.

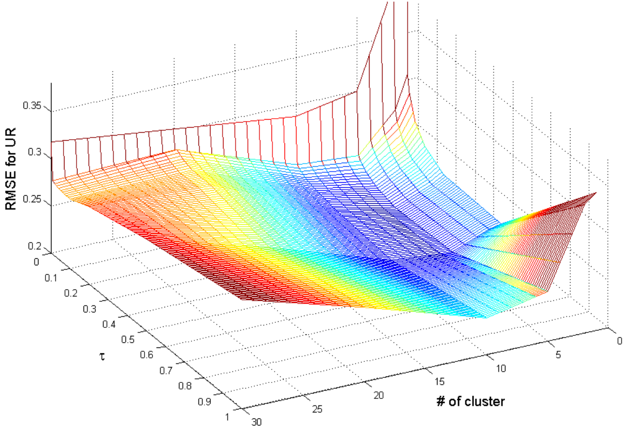
Choosing of Parameters

- Number of slices H;
  - It's not crucial, there are theoretical results (Li, 2000) indicating the SIR outputs do not change much for a wide range of H.
  - Choose $H = 10$.
- Number of lags l;
  - It's not crucial, usually all the lags will be selected into one cluster.
  - To be consistent to Stock and Watson (2005) paper, choose $l = 4$.
- Number of clusters c;
- Shrinkage parameter for regularization τ.
  - Choose their values based on the RMSE.

UIC University of Illinois at Chicago

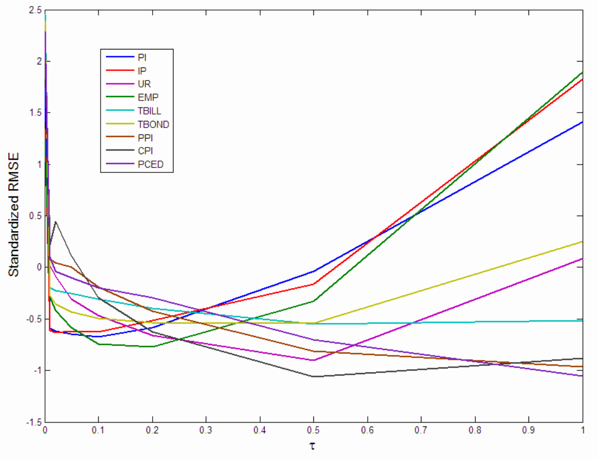
Choosing c and τ

RMSE of Unemployment Rate for different values of c and τ



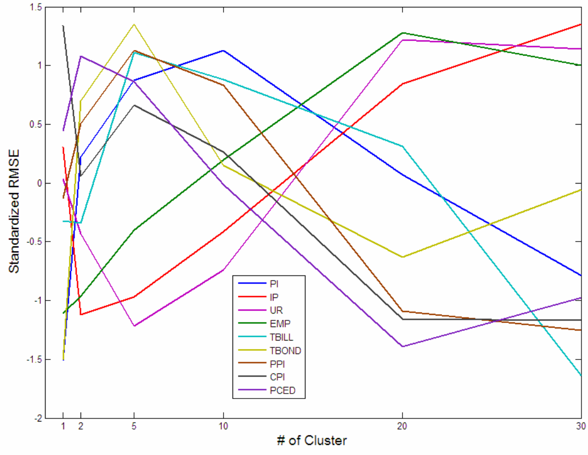
Choosing c and τ

Standardize RMSE vs. τ when $c = 10$



Choosing c and τ

Standardize RMSE vs. c when $\tau = 0.1$



University of Illinois at Chicago

Results

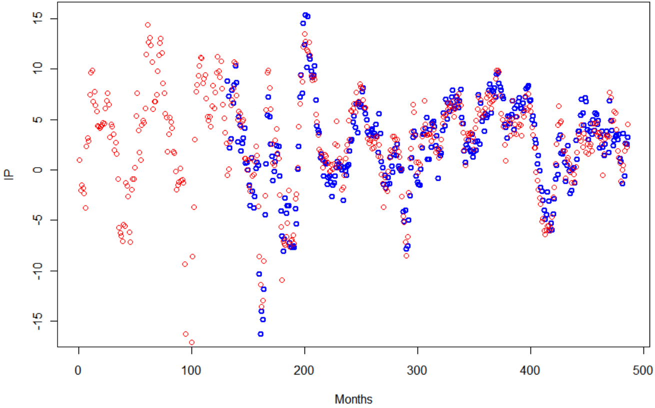
When $c = 10$ and $\tau = 0.1$

	Cluster-based Regularized SIR	Best Case Stock & Watson	Worst Case Stock & Watson
PI	**1.71**	2.84	7.56
IP	**1.91**	4.11	10.96
UR	**0.23**	0.42	0.87
EMP	**0.79**	1.48	3.39
TBILL	**0.99**	1.31	2.24
TBOND	**0.63**	1.02	1.56
PPI	**6.27**	**3.04**	9.46
CPI	**2.80**	**1.44**	3.97
PCED	**2.06**	**1.15**	3.20

UIC University of Illinois at Chicago

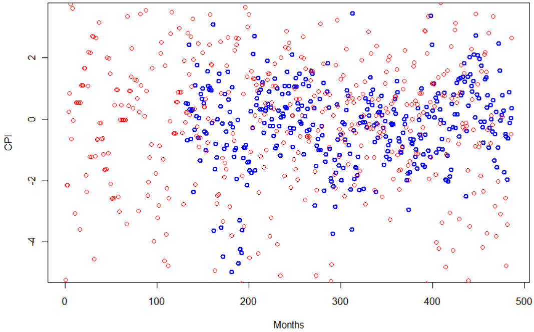
Results

IP, Red: Original Value; Blue: Fitted Value ($c = 10$, $\tau = 0.1$)



Drawbacks of SIR

- The inverse regression method is to detect the variation of $E(x|Y = y)$, if y has no tendency to the change of the other variables, the inverse regression method may not work.



Further Works

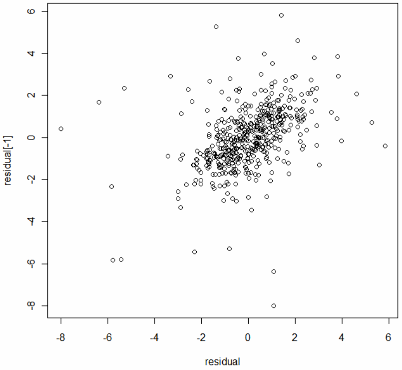
- Find a better model instead of linear model to fit

$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon).$$

Further Works

- Find a better model instead of linear model to fit

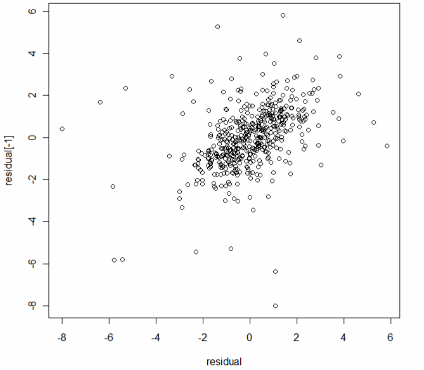
$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \ldots, \beta_K'\mathbf{x}, \epsilon).$$



Further Works

- Find a better model instead of linear model to fit

$$y = g(\beta_1'\mathbf{x}, \beta_2'\mathbf{x}, \dots, \beta_K'\mathbf{x}, \epsilon).$$



- Missing value problem.

References

- Ker-Chau Li, Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, Vol. 86, No. 414, pp. 316–327, 1991.
- N. Duan and K.C. Li., Slicing regression: a link-free regression method. *The Annals of Statistics*, Vol. 19, No. 2, pp. 505–530, 1991.
- Stock, J.H. and Watson, M.W., An empirical comparison of methods for forecasting using many predictors, *Manuscript, Princeton University*, 2005.
- Zhihong Chen, Azhar Iqbal and Huiwen Lai, Forecasting the Probability of US Recessions: A Probit and Dynamic Factor Modeling Approach, *Canadian Journal of Economics*, 2010.
- Eaton, M.L., A characterization of spherical distributions, *Journal of Multivariate Analysis*, Vol. 20, No. 2, pp. 272–276, 1986.
- Friedman, J.H., Regularized discriminant analysis, *Journal of the American statistical association*, Vol. 64, No. 405, pp. 165–175, 1989.
- Scrucca, L., Regularized Sliced Inverse Regression with applications in classification, *Data Analysis, Classification and the Forward Search*, pp. 59–66, 2006.
- Becker, C. and Fried, R., Sliced inverse regression for high-dimensional time series, *Exploratory data analysis in empirical research: proceedings of the 25th Annual Conference of the Gesellschaft f "ur Klassifikation eV*, University of Munich, March 14-16, 2001.
- Zhong, W. and Zeng, P. and Ma, P. and Liu, J.S. and Zhu, Y., RSIR: regularized sliced inverse regression for motif discovery, *Bioinformatics*, Vol. 21, No. 22, pp. 4169–4175, 2005.
- Kuentz, V. and Saracco, J., Cluster-based Sliced Inverse Regression, *Journal of the Korean Statistical Society*, 2009.
- Ker-Chau Li, High dimensional data analysis via the SIR/PHD approach, *Manuscript*, 2000.

Q & A

Yue Yu
Oct. 6, 2010