Lecture 14: Markov Neighborhood Regression for High-Dimensional Inference

High-Dimensional Data Research

- Variable Selection:
  - Frequentist (regularization) methods: Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001); MCP (Zhang, 2010), rLasso (Song and Liang, 2015);
  - Bayesian methods: Subset Modeling (Liang et al., 2013), Split-and-Merge (Song and Liang, 2015).
- Sure Independence Screening: Fan and Lv (2008), Fan and Song (2010)
- Graphical models:
  - nodewise regression (Meinshausen and Buhlmann, 2006)
  - graphical Lasso (Yuan and Lin, 2007; Friedman et al., 2008)
  - ψ-learning (Liang et al., 2015)

High-Dimensional Inference

Consider a high-dimensional linear regression:

$$Y = \boldsymbol{X}\beta + \epsilon,$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, and the sample size n is much smaller than the dimension p (small-n-large-p).

We are interested in assessing uncertainty of the model, in particular, constructing the confidence interval for each regression coefficient β_i and calculating the associated p-values.

Later, you will see that the proposed method also works for generalized linear models and improves accuracy of variable selection.

High-Dimensional Inference

- desparsified Lasso (van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014)

  $$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \hat{\Theta}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta})/n,$$

  where $\hat{\boldsymbol{\beta}}$ is the original Lasso estimator, and $\hat{\Theta}$ is an approximation for the inverse of $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$.

  $$\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) = \hat{\Theta}\boldsymbol{X}^T\boldsymbol{\epsilon}/\sqrt{n}+\sqrt{n}(I_p-\hat{\Theta}\hat{\boldsymbol{\Sigma}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \sim N(0, \hat{\sigma}^2\hat{\Theta}\hat{\boldsymbol{\Sigma}}\hat{\Theta}^T),$$

  where I_p denotes the $p \times p$ identity matrix.

High-Dimensional Inference

- **Multi sample-splitting**: Splitting the samples into half and half, using the first half for variable selection and the second half with the reduced set of selected variables for statistical inference in terms of *p*-values; repeating this process for many times; and aggregating the *p*-values obtained in the process for statistical inference.
- **Ridge projection**: It can be viewed as a direct extension of the low-dimensional ridge regression.
- **residual-type bootstrapping**: supper-efficiency phenomenon
- **Other methods**: covariance test, group-bound

Graph Theory

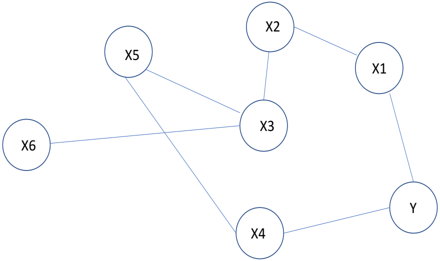


Figure: An illustrative graphical model

Graph Theory

An undirected graph is a pair $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is the set of vertices and $\boldsymbol{E} = (e_{ij})$ is the adjacency matrix.

- If two vertices $i, j \in \boldsymbol{V}$ forms an edge then we say that i and j are *adjacent* and set $e_{ij} = 1$.
- A *path* of length $l > 0$ from v_0 to v_l is a sequence $v_0, v_1, \ldots, v_l$ of distinct vertices such that $e_{v_{k-1},v_k} = 1$ for all $k = 1, \ldots, l$.
- The subset $\boldsymbol{U} \subset \boldsymbol{V}$ is said to *separate* $\boldsymbol{I} \subset \boldsymbol{V}$ from $\boldsymbol{J} \subset \boldsymbol{V}$ if for every $i \in \boldsymbol{I}$ and $j \in \boldsymbol{J}$, all paths from i to j have at least one vertex in $\boldsymbol{U}$.

Graph Theory

- $P_{\boldsymbol{V}}$ is said to satisfy the *Markov property* with respect to $\boldsymbol{G}$ if for every triple of disjoint sets $\boldsymbol{I}, \boldsymbol{J}, \boldsymbol{U} \subset \boldsymbol{V}$, it holds that $X_{\boldsymbol{I}} \perp X_{\boldsymbol{J}} | X_{\boldsymbol{U}}$ whenever $\boldsymbol{U}$ separates $\boldsymbol{I}$ and $\boldsymbol{J}$ in $\boldsymbol{G}$.
- Let $\xi_j = \{k : e_{jk} = 1\}$ denote the neighboring set of X_j in $\boldsymbol{G}$. Following from the Markov property of the Gaussain graphical model (GGM), we have $X_j \perp X_i | \boldsymbol{X}_{\xi_j}$ for any $i \in \boldsymbol{V} \setminus \xi_j$, as ξ_j forms a separator between X_i and X_j.
- For convenience, we call ξ_j the minimum Markov neighborhood of X_j in $\boldsymbol{G}$, and call any subset $\boldsymbol{s}_j \supset \xi_j$ a Markov neighborhood of X_j in $\boldsymbol{G}$.

A Simple Mathematical fact

- Let $\boldsymbol{S}_1 = \{2, \ldots, d\}$ denote a Markov neighborhood of X_1, let Σ_d denote the covariance matrix of $\{X_1\} \cup \boldsymbol{X}_{\boldsymbol{S}_1}$, and partition Θ as
$$\Theta = \begin{bmatrix} \Theta_d & \Theta_{d,p-d} \\ \Theta_{p-d,d} & \Theta_{p-d} \end{bmatrix}, \qquad (1)$$
where the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ are exactly zero, as $X_1 \perp \boldsymbol{X}_{\boldsymbol{V} \setminus (\{1\} \cup \boldsymbol{S}_1)} | \boldsymbol{X}_{\boldsymbol{S}_1}$ holds.
- Inverting the partitioned matrix, we have $\Sigma_d = (\Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d})^{-1}$, and
$$\Sigma_d^{-1} = \Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}. \qquad (2)$$
- Since the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ are exactly zero, the $(1,1)$th element of $\Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}$ is exactly zero.

A Simple Mathematical fact

- *If we assume that* $\{X_1\} \cup \boldsymbol{X}_{\boldsymbol{S}_1} \supset \boldsymbol{X}_{\boldsymbol{S}_*}$*, where* $\boldsymbol{S}_*$ *denote the set of true predictors, then the statistical inference for* β_1 *from the original model will be exactly the same as that from the subset regression*

  $$Y = \beta_0' + X_1\beta_1 + X_2\beta_2' + \ldots + X_d\beta_d' + \epsilon, \tag{3}$$

  where the prime on β_i's for $i \neq 1$ indicates that those regression coefficients might be modified by the subset regression.

- Since $\boldsymbol{S}_1$ forms a Markov neighborhood of X_1, we call (3) a Markov neighborhood regression, which reduces the high-dimensional inference problem to a series of low-dimensional inference problems.

Markov Neighborhood Regression: Variable selection-based

(a) (*Graphical Model Construction*) Construct a GGM for $\boldsymbol{X}$ and obtain a consistent estimate of the minimum Markov neighborhood for each variable. Denote the estimates by $\hat{\xi}_j$ for $j = 1, 2, \ldots, p$.

(b) (Variable selection) Conduct variable selection for the model to get a consistent estimate of $\boldsymbol{S}_*$, the set of true predictors. Denote the estimate by $\hat{\boldsymbol{S}}_*$.

(c) (Subset regression) For each variable X_j, $j = 1, \ldots, p$, let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{S}}_*$ and run an Ordinary Least Square (OLS) regression with the predictors given by $\boldsymbol{X}_{D_j}$, i.e.,

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X}_{D_j}\boldsymbol{\beta}_{D_j} + \epsilon, \qquad (4)$$

where $\epsilon \sim N(0, \sigma^2 I_n)$ and I_n is an $n \times n$-identity matrix. Conduct inference for β_j, including the estimate, confidence interval and p-value, based on the output of (4).

MNR: Justification

Lemma 1 Let $\hat{\xi}_j \supseteq \xi_j$ denote any Markov neighborhood of feature $x^{(j)}$, let $\hat{\boldsymbol{S}}_* \supseteq \boldsymbol{S}_*$ denote any reduced feature space, and let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{S}}_*$. Consider the subset regression (4). Let $\hat{\boldsymbol{\beta}}_{D_j}$ denote the OLS estimator of $\boldsymbol{\beta}_{D_j}$ from the subset regression, and let $\hat{\beta}_j$ denote the element of $\hat{\boldsymbol{\beta}}_{D_j}$ corresponding to the variable X_j. If $|D_j| = o(n^{1/2})$, as $n \to \infty$, the following results hold:

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j^*) \sim N(0, \sigma^2\theta_{jj})$, where θ_{jj} is the (j,j)-th entry of the precsion matrix Θ.

(ii) $\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{n\hat{\sigma}_n^2\hat{\theta}_{jj}}} \sim N(0,1)$, where $\hat{\sigma}_n^2 = (\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})^T(\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})/(n-d-1)$, $\hat{\theta}_{jj}$ is the (j,j)-th entry of the matrix $\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{D_j}^{(i)}(\boldsymbol{x}_{D_j}^{(i)})^T\right]^{-1}$, and $\boldsymbol{x}_{D_j}^{(i)}$ denotes the i-th row of $\boldsymbol{X}_{D_j}$.

MNR: Remark

Remark: Lemma 1 assumes that $\hat{\boldsymbol{S}}_* \supseteq \boldsymbol{S}_*$ and $|D_j| = o(n^{1/2})$. For the case that n is finite, we have $(n - |D_j| - 1)\hat{\sigma}_n^2/\sigma^2 \sim \chi^2(n - |D_j| - 1)$, independent of $\hat{\boldsymbol{\beta}}_{D_j}$, by the standard theory of OLS estimation. Therefore, we can use $t(n - |D_j| - 1)$ to approximate the distribution of $\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{n\hat{\sigma}_n^2\hat{\theta}_{jj}}}$; that is, *the estimate, p-value and confidence interval of β_j can be calculated from (4) as in conventional low-dimensional multiple linear regression.*

MNR: Justification

Theorem 1 (Validity of Algorithm 1) If the conditions (A0)-(A9) hold, the ψ-learning algorithm is used for GGM construction in step (a), and the SCAD algorithm is used for variable selection in step (b), then for each $j \in \{1, 2, \dots, p_n\}$, we have $\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{n\hat{\sigma}_n^2 \hat{\theta}_{jj}}} \sim N(0,1)$ as $n \to \infty$, where $\hat{\sigma}_n^2 = (\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})^T(\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})/(n-d-1)$, $\hat{\theta}_{jj}$ is the (j,j)-th entry of the matrix $\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_{D_j}^{(i)}(\boldsymbol{x}_{D_j}^{(i)})^T\right]^{-1}$, and $\boldsymbol{x}_{D_j}^{(i)}$ denotes the i-th row of $\boldsymbol{X}_{D_j}$.

Markov Neighborhood Regression: Variable screening-based

1 (Correlation screening) Apply the correlation screening procedure to $\boldsymbol{X}$ to obtain a reduced neighborhood $\hat{\xi}_j \subseteq \{1, \ldots, p\}$ for each feature x_j.

2 (Variable screening) Apply a sure independence screening procedure, with Y as the response variable and $\boldsymbol{X}$ as predictors, to obtain a reduced feature set, $\hat{\boldsymbol{S}}_* \subseteq \{1, \ldots, p\}$, with the size $|\hat{\boldsymbol{S}}_*| = O(\sqrt{n}/\log(n))$.

3 (Subset Regression) For each variable X_j, $j = 1, \ldots, p$, run the OLS regression with the predictors given by $\{X_j\} \cup \boldsymbol{X}_{\hat{\xi}_j} \cup \boldsymbol{X}_{\hat{\boldsymbol{S}}_*}$, i.e., the subset regression (4). Conduct inference for β_j, including the estimate, confidence interval and p-value, based on the output of the subset regression.

MNR: Justification

Theorem 2 If the conditions (A0), (A9), and (B1)-(B4) (given in the Appendix) hold, then for each $j \in \{1, 2, \ldots, p_n\}$, we have $\frac{\hat{\beta}_j - \beta_j^*}{\sqrt{n\hat{\sigma}_n^2 \hat{\theta}_{jj}}} \sim N(0, 1)$ as $n \to \infty$, where $\hat{\sigma}_n^2 = (\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})^T(\boldsymbol{y} - \boldsymbol{x}_{D_j}\hat{\boldsymbol{\beta}}_{D_j})/(n - d - 1)$, $\hat{\theta}_{jj}$ is the (j, j)-th entry of the matrix $\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{D_j}^{(i)}(\boldsymbol{x}_{D_j}^{(i)})^T\right]^{-1}$, and $\boldsymbol{x}_{D_j}^{(i)}$ denotes the i-th row of $\boldsymbol{X}_{D_j}$.

Variable Selection versus Variable Screening

- Compared to the variable selection-based algorithm, the variable screening-based algorithm is faster and more robust.
- The screening-based algorithm reduces the risk of losing variables in the true minimum Markov neighborhood as well as the risk of losing true predictors, while maintaining the validity of MNR.
- Since more variables will be included in the subset regression, the resulting confidence interval will be slightly wider.

Generalized Linear Models (GLMs)

Lemma 2 Let $\hat{\xi}_j \supseteq \xi_j$ denote any Markov neighborhood of feature x_j, let $\hat{\boldsymbol{S}}_* \supseteq \boldsymbol{S}_*$ denote any reduced feature space, and let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{\boldsymbol{S}}_*$. Consider a subset GLM with the predictorts $\boldsymbol{X}_{D_j}$, let $\hat{\boldsymbol{\beta}}_{D_j}$ denote the MLE of $\boldsymbol{\beta}_{D_j}$, and let $\hat{\beta}_j$ denote the component of $\hat{\boldsymbol{\beta}}_{D_j}$ corresponding to feature X_j. If $|D_j| = o(n^{1/2})$, then, as $n \to \infty$, the following results hold:

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j^*) \sim N(0, k_{jj})$, where k_{jj} denotes the (j,j)-th entry of the inverse of the Fisher information matrix $K = I^{-1} = [E(b''(\boldsymbol{x}^T \boldsymbol{\beta}^*)\boldsymbol{x}\boldsymbol{x}^T)]^{-1}$, and $\boldsymbol{\beta}^*$ denotes the true regression coefficients.

(ii) $\sqrt{n}(\hat{\beta}_j - \beta_j^*)/\sqrt{\hat{k}_{jj}} \sim N(0,1)$, where $\hat{k}_{jj}$ denotes the (j,j)-th entry of the inverse of the observed information matrix $J_n(\hat{\boldsymbol{\beta}}_{D_j}) = -\sum_{i=1}^n H_{\hat{\boldsymbol{\beta}}_{D_j}}(\log f(y_i|\boldsymbol{\beta}_{D_j}, \boldsymbol{x}_{D_j}))/n$ and $H_{\hat{\boldsymbol{\beta}}_{D_j}}(\cdot)$ denotes the Hessian matrix evaluated at the MLE $\hat{\boldsymbol{\beta}}_{D_j}$.

Joint Inference

- Let $\boldsymbol{A} \subset \boldsymbol{V}$ denote a subset of predictors for which the joint inference is desired. Define $\xi_{\boldsymbol{A}} = \cup_{j \in \boldsymbol{A}} \xi_j$ as the union of the minimum Markov neighborhoods of the variables in $\boldsymbol{A}$. Let $\boldsymbol{M} = \boldsymbol{A} \cup \hat{\xi}_A \cup \hat{\boldsymbol{S}}_*$. Then a subset regression can be conducted with the predictors included in $\boldsymbol{M}$.
- For high-dimensional linear regression, if $|\boldsymbol{A}| = O(1)$, then, similar to Theorem 1, we can show $\sqrt{n}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \sim N(0, \sigma^2 \Theta_{AA})$, where Θ_{AA} denotes the submatrix of the precision matrix Θ constructed by its A rows and A columns.
- For high-dimensional GLMs, we have $\sqrt{n}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*) \sim N(0, K_{AA})$, where K_{AA} denotes the submatrix of $K = [E(b''(\boldsymbol{x}^T \boldsymbol{\beta}^*)\boldsymbol{x}\boldsymbol{x}^T]^{-1}$ constructed by its A rows and A columns.

Example 1

We first generated a dataset from a linear regression with $n = 2000$ and p=50, where σ^2 was set to 1, the covariates $\boldsymbol{X}$ were generated from a multivariate Gaussian distribution with mean $\mathbf{0}$ and a Toeplitz covariance matrix $\Sigma_{i,j} = 0.9^{|i-j|}$ for $i, j = 1, \ldots, p$, and the true regression coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_5) = (1, 0.2, 0.4, -0.3, -0.5, 1.0)$ and $\beta_6 = cdots = \beta_p = 0$.

Example 1: Confidence Interval

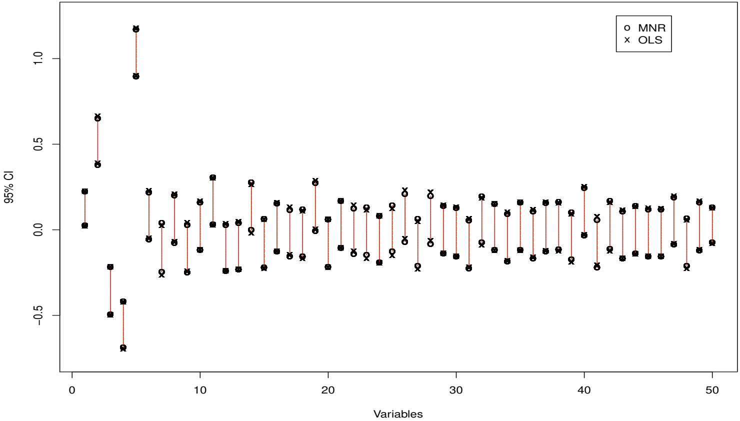


Figure: Comparison of confidence intervals of $\beta_1, \ldots, \beta_p$ produced by ordinary least square (red) and Markov neighborhood regression (black) for a dataset with $n = 2000$ and $p = 50$.

Example 2

We generated 100 datasets from a linear regression with $n = 200$ and p=500, where σ^2 was set to 1, the covariates $\boldsymbol{X}$ were generated from a multivariate Gaussian distribution with mean $\mathbf{0}$ and a Toeplitz covariance matrix $\Sigma_{i,j} = 0.9^{|i-j|}$ for $i, j = 1, \ldots, p$, and the true regression coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_5)$= $(1, 2, 4, -3, -5, 10)$ and $\beta_6 = \cdots = \beta_p = 0$.

Example 2

Table: Coverage rates and widths of the 95% confidence intervals for the Toeplitz-covariance linear regression model.

Measure		Desparsified-Lasso	Ridge	MNR
Coverage	signal	0.384(0.049)	0.576(0.049)	0.956(0.021)
	noise	0.965(0.018)	0.990(0.010)	0.950(0.022)
Width	signal	0.673(0.005)	1.086(0.010)	0.822(0.011)
	noise	0.691(0.005)	1.143(0.008)	0.869(0.007)

Example 2: Screening Algorithm

Table: Coverage rates and widths of the 95% confidence intervals produced by the Screening MNR Algorithm for the Toeplitz-covariance linear regression with different values of m, which controls the size of Markov neighborhoods.

Measure		$m = 3$	$m = 5$	$m = 8$	$m = 15$	$m = 20$
Coverage	signal	0.956(0.021)	0.962(0.019)	0.956(0.021)	0.958(0.020)	0.954(0.021)
	noise	0.952(0.021)	0.950(0.022)	0.951(0.022)	0.951(0.022)	0.949(0.022)
Width	signal	1.023(0.035)	0.854(0.014)	0.839(0.011)	0.857(0.011)	0.876(0.011)
	noise	2.566(0.019)	1.610(0.054)	0.902(0.008)	0.935(0.007)	0.963(0.008)

Example 2: MNR for Variable selection

- Convert p-values produced by the subset regressions to z-scores using the transformation:

$$Z_i^{(j)} = \Phi^{-1}(1 - q_i^{(j)}), \quad i = 1, \dots, p, \quad j = 1, 2, \dots, n.$$

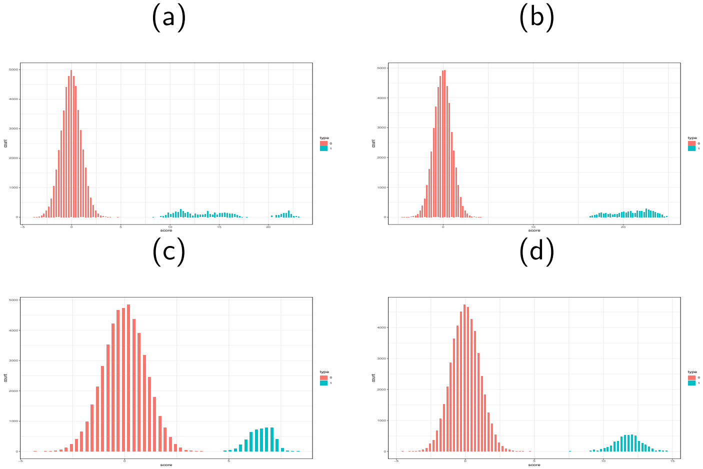
- Conduct multiple hypothesis tests.

Example 2: MNR for Variable selection

Table: Variable selection for the Toeplitz-covariance linear regression with the MNR, SIS-SCAD, SIS-MCP and SIS-Lasso methods.

Measure	MNR			SIS-SCAD	SIS-MCP	SIS-Lasso
	$q = 0.0001$	$q = 0.001$	$q = 0.01$			
FSR	0	0.004	0.022	0.127	0.175	0.819
NSR	0	0	0	0	0	0

Example 2: MNR for Variable selection


(a)

(b)

(c)

(d)

Figure: Histograms of z-scores of the features produced by MNR for (a) Toeplitz-covariance linear regression, (b) AR(2)-precision linear regression, (c) AR(2)-precision logistic regression, (d) AR(2)-precision Cox regression.

Example 2: MNR for joint inference

Table: Coverage rates of the 95% Bonferroni joint confidence intervals produced by MNR for sets of parameters.

Parameters	(β_1, β_2)	$(\beta_3, \beta_4, \beta_5)$	(β_1, β_6)	(β_7, β_{10})	$(\beta_{20}, \beta_{200}, \beta_{400})$
Joint coverage rate	0.97(0.017)	0.95(0.022)	0.93(0.026)	0.97(0.017)	0.93(0.026)

Example 3

Consider the AR(2)-precision matrix $\Theta = (\theta_{ij})$ given by

$$\theta_{ij} = \begin{cases} 0.5, & \text{if } |j-i| = 1, i = 2, ..., (p-1), \\ 0.25, & \text{if } |j-i| = 2, i = 3, ..., (p-2), \\ 1, & \text{if } i = j, i = 1, ..., p, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

- Linear: 100 datasets were generated from the linear regression with $n = 200$, $p = 500$, $\sigma^{=}1$, the features generated from a zero-mean Gaussian with an AR(2)-precision matrix, $(\beta_0, \cdots, \beta_5) = (1, 2, 2.5, 3, 3.5, 4)$ and all others equal to 0.
- Logistic: same setting as linear but with $n = 300$.
- Cox: same setting as linear but with $n = 300$ and $\beta_1 = \cdots = \beta_5 = 1$ and all others equal to 0.

Example 3

Table: Coverage rates of 95% confidence intervals for the AR(2)-precision linear ($n = 200$, $p = 500$), logistic ($n = 300$,$p = 500$) and Cox regression ($n = 300$, $p = 500$).

Regression		Desparsified-Lasso	Ridge	MNR
Linear	signal	0.2300(0.0421)	0.3340(0.0447)	**0.9500(0.0218)**
	noise	0.9640(0.0186)	0.9922(0.0088)	**0.9503(0.0217)**
Logistic	signal	0.004(0.0063)	0(0)	**0.9320(0.0252)**
	noise	0.9953(0.0068)	1.0(4.5e-4)	**0.9373(0.0242)**
Cox	signal	—	—	**0.9140(0.0281)**
	noise	—	—	**0.9354(0.0246)**

Example 3

Table: Widths of 95% confidence intervals for the AR(2)-precision linear ($n = 200$, $p = 500$), logistic ($n = 300$,$p = 500$) and Cox regression ($n = 300$, $p = 500$).

Regression		Desparsified-Lasso	Ridge	MNR
Linear	signal	0.2810(0.0027)	0.4481(0.0043)	0.2806(0.0022)
	noise	0.2723(0.0024)	0.4335(0.0036)	0.2814(0.0024)
Logistic	signal	0.6424(0.0101)	1.0775(0.0110)	1.9473(0.0529)
	noise	0.5782(0.0081)	1.0100(0.0095)	0.9799(0.0132)
Cox	signal	—	—	0.3356(0.0018)
	noise	—	—	0.2683(0.0017)

Bias of Desparsified-Lasso Estimates

Let Z_j denote the residual of the regression X_j versus all other features $\boldsymbol{X}[-j]$, and let $P_{jk} = X_k^T Z_j / X_j^T Z_j$. Then the following identity holds

$$\frac{Y' Z_j}{X_j^T Z_j} = \beta_j + \sum_{k \neq j} P_{jk} \beta_k + \frac{\epsilon' Z_j}{X_j^T Z_j}. \tag{6}$$

Plugging the Lasso estimator $\hat{\boldsymbol{\beta}}_{Lasso}$ (of the regression Y versus $\boldsymbol{X}$) into (6) leads to the bias-corrected estimator

$$\hat{\beta}_{bc,j} = \frac{Y' Z_j}{X_j^T Z_j} - \sum_{k \neq j} P_{jk} \hat{\beta}_{Lasso,k} = \hat{\beta}_{Lasso,j} + Z_j'(Y - \boldsymbol{X} \hat{\boldsymbol{\beta}}_{Lasso}) / Z_j' X_j, \tag{7}$$

for $j = 1, 2, \ldots, p$. The bias-corrected estimator is still biased if the sample size n is not sufficiently large.

Bias of Desparsified-Lasso Estimates

Table: Regression coefficient estimates (averaged over 100 independent datasets) produced by MNR and desparsified-Lasso for the AR(2)-precision linear regression (with $|\boldsymbol{S}_*| = 5$, $p = 500$ and $n = 200$).

Method	Measure	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
—	true	2	2.5	3	3.5	4	0	0	0
desparsified	$\hat{\boldsymbol{\beta}}_{bc}$	1.841	2.274	2.698	3.270	3.849	-0.051	-0.007	0.016
	SD	(0.008)	(0.009)	(0.009)	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
MNR	$\hat{\boldsymbol{\beta}}_{MNR}$	1.997	2.503	2.994	3.498	4.001	0.014	0.004	-0.002
	SD	(0.006)	(0.008)	(0.008)	(0.007)	(0.006)	(0.007)	(0.008)	(0.008)

Causal Structure Discovery

- The causal relationship for a pair or more variables refers to a *persistent association* which is expected to exist in all situations without being affected by the values of other variables.
- In statistics, the causal relationship or the *persistent association* can be determined using conditional independence tests. For a large set of variables, a pair of variables are considered to have no direct causal relationship if a subset of the remaining variables can be found such that conditioning on this subset of variables, the two variables are independent.
- PC algorithm (Spirtes et al., 2000), PC-simple (Buhlmann et al., 2010)

Simplied MNR Algorithm

(a) (Variable screening) Apply a sure independence screening procedure with Y as the response variable and $\boldsymbol{X}$ as predictors, to obtain a reduced feature set, $\hat{\boldsymbol{S}}_* \subseteq \{1, \ldots, p\}$, with the size $|\hat{\boldsymbol{S}}_*| = O(\sqrt{n}/\log(n))$.

(b) (Minimum Markov neighborhood determination) For each variable $X_j \in \hat{\boldsymbol{S}}_*$, apply a sure independence screening procedure to obtain a reduced neighborhood $\hat{\xi}_j \subseteq \{1, \ldots, p\}$.

(c) (Subset Regression) For each variable $X_j \in \hat{\boldsymbol{S}}_*$, run a subset regression with the predictors given by $\{X_j\} \cup \boldsymbol{X}_{\hat{\xi}_j} \cup \boldsymbol{X}_{\hat{\boldsymbol{S}}_*}$. Conduct inference for β_j, including the estimate, confidence interval and p-value, based on the output of the subset regression.

(d) (Causal Structure Discovery) Conduct a multiple hypothesis test to identify causal predictors based on the p-values calculated in step (c).

Causal Structure Discovery

- Linear Regression: Gaussian graphical model, faithfulness is required
- Logistic Regression: Faithfulness to the mixed graphical model
- Cox Regression: Bayesian network (as Y is non-Gaussian, non-multinomial, and with missing observations)

CCLE Drug Response

The cancer cell line encyclopedia (CCLE) database, which is publicly available at *www.broadinstitute.org/ccle*. The dataset consisted of 8-point dose-response curves for 24 chemical compounds across over 400 cell lines. For different chemical compounds, the numbers of cell lines are slightly different. For each cell line, it consisted of the expression value of $p = 18,988$ genes. We used the area under the dose-response curve, which is termed as activity area, to measure the sensitivity of a drug for each cell line.

Compared to other measurements, such as IC_{50} and EC_{50}, the activity area could capture the efficacy and potency of the drug simultaneously.

CCLE Drug Response

Table: Comparison of drug sensitive genes selected by desparsified Lasso, ridge projection, multi-split and MNR for 24 anti-cancer drugs, where * indicates that this gene was significantly selected and the number in the parentheses denotes the width of the 95% confidence intervals produced by the method.

Drug	Des-Lasso	Ridge	Multi-Split	MNR
17-AAG	–	–	NQO1*(0.138)	NQO1*(0.088)
AEW541	–	F3(0.076)	SP1(0.176)	SLC10A7(0.116)
AZD0530	–	PPY2(0.966)	SYN3(0.705)	BEST3(0.152)
AZD6244	–	OSBPL3(0.161)	SPRY2*(0.084) LYZ*(0.069) RNF125*(0.084)	LYZ*(0.041) RNF125*(0.051)
Erlotinib	–	LRRN1(0.102)	PCDHGC3(0.684)	PRRG4(0.043)
Irinotecan	–	SLFN11(0.091)	ARHGAP19*(0.134) SLFN11*(0.044)	SLFN11*(0.032)

Identification of Cancer Driver Gene

The Lymph dataset consists of $n = 148$ samples with 100 node-negative cases (low risk for breast cancer) and 48 node-positive cases (high risk for breast cancer) as our binary response. For each sample, there are $p = 4512$ genes that showed evidence of variation above the noise level for futher study.

Identification of Cancer Driver Gene

Table: Comparison of the cancer driver gene selected by the MNR, desparsified Lasso and ridge projection methods for the Lymph dataset, where * indicates that this gene was significantly selected and the number in the parentheses denotes the width of the 95% confidence intervals produced by the method.

	Desparsified Lasso	Ridge	MNR
Gene	RGS3	RGS3	RGS3*
CI	(1.145,5.748)	(-0.251,2.249)	(2.651,5.999)
Width	4.603	2.500	3.348

Discussion

- We have proposed an innovative method for conducting statistical inference, assessing p-values and constructing confidence intervals, for high-dimensional regression and generalized linear models.
- The proposed method can be very fast compared to the existing methods: embrassingly parallel structure
- The proposed method can deal with ultra-high dimensional problems, as it has reduced the problem to a series of low-dimensional problems.
- Markov neighbodhood, or conditional independence set, is a useful concept and can be used in many problems.