Contrasted Penalized Integrative Analysis

Shuangge Ma

School of Public Health, Yale University

DIMACS Workshop on Statistical Analysis of Network Dynamics and Interactions, Nov 7-8, 2013. Rutgers

The High Dimensional Era of Statistics

One of the biggest buzzwords of this year: big data.

One type of big data: large p, small n.

Such data have been encountered in medicine, finance, engineering, and even social science.

Our Analysis Strategy

Consider a generic model $Y \sim \phi(\beta' X)$ where β is the unknown regression coefficients. Assume a sparse or sparsified model.

With n iid observations, denote $R(\beta)$ as the loss function.

Analysis goal: from a large number of candidate covariates, identify a few that are associated with the response & estimate the unknown parameters.

Penalization: a generic estimation and variable selection technique

$$\widehat{\beta} = argmin_{\beta}\{R(\beta) + P(\beta)\}$$

$P(\beta)$ is the model complexity measure. It often has the form $P(\beta) = \lambda \times \sum_{j=1}^{p} f(|\beta_j|)$, that is, it is separable.

Covariates that correspond to the nonzero components of $\widehat{\beta}$ are identified as associated with the response.

When $\log(p)/n \to 0$ plus a few other mild conditions, $Pr(sign(\widehat{\beta}) = sign(\beta)) \to 1$.

Integrative Analysis

Important covariates identified from the analysis of high-dimensional datasets often have low reproducibility.

There are many contributing factors, among which is the small sample sizes of individual studies.

If concerned with sample size, let's have more samples (for example, NCI consortiums).

Multi-dataset approaches: meta-analysis and integrative analysis

Assume M independent studies, and n^m iid observations in study $m(= 1, \ldots, M)$. $\sum_m n^m << p$.

In study m, denote Y^m as the response variable and X^m as the length p covariates. Assume $Y^m \sim \phi(\boldsymbol{\beta}^{m\prime} X^m)$. Denote $R^m(\boldsymbol{\beta}^m)$ as the objective function, for example the negative log-likelihood function.

The overall objective function $R(\boldsymbol{\beta}) = \sum_m R^m(\boldsymbol{\beta}^m)$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M)$.

Denote β_j^m as the $j$the component of $\boldsymbol{\beta}^m$. Denote $\boldsymbol{\beta}_j = (\beta_j^1, \ldots, \beta_j^M)'$, which represents the effects of covariate j across M datasets.

Two Candidate Models

Homogeneity model all datasets share the same set of susceptibility covariates. That is, $\boldsymbol{\beta}^m$'s have the same sparsity structure.

Heterogeneity model a covariate can be associated with outcome in some datasets but not others. It includes the homogeneity model as a special case and is more flexible.

Penalized marker selection

Consider the penalized estimate

$$\widehat{\boldsymbol{\beta}} = argmin\left\{R(\boldsymbol{\beta}) + P_{\lambda,\gamma}(\boldsymbol{\beta})\right\}.$$

Our working penalty is MCP $\rho(t;\lambda_1,\gamma) = \lambda_1 \int_0^{|t|} \left(1 - \frac{x}{\lambda_1\gamma}\right)_+ dx$ proposed by Dr. Cunhui Zhang.

Under the homogeneity model

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho\left(||\boldsymbol{\beta}_j||_2; \sqrt{M_j}\lambda_1, \gamma\right)$$

which conducts one-dimensional selection. M_j is the "size" of $\boldsymbol{\beta}_j$. When the M datasets have matched covariate sets, $M_j \equiv M$.

Under the heterogeneity model

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho\left(||\boldsymbol{\beta}_j||_1; \sqrt{M_j}\lambda_1, \gamma\right)$$

which conducts two-dimensional selection. The $||\cdot||_1$ norm can be replaced with for example MCP, leading to a composite MCP penalty.

The above penalization approaches account for the grouping structure of regression coefficients.

However, there exist other structures of covariates (and regression coefficients) that have not been effectively accounted for.

Here we consider two specific examples: a within-dataset structure and an across-dataset structure.

Within-dataset structure

Here the structure describes the interplay of covariates within the same dataset.

Network based Analysis

A node corresponds to a covariate.

The most important characteristic of a network is the adjacency measure, which quantifies how closely two nodes are connected. The adjacency measure is often defined based on the notion of similarity between nodes.

Consider a_{jk}, which measures the strength of connection between nodes (covariates) j and k.

Assume undirected network where $a_{jk} = a_{kj}$ for $j, k = 1, \ldots, p$.

Construction of adjacency matrix

Denote r_{jk} as the Pearson's correlation coefficient, and π_{jk} as the canonical correlation between covariate j and k.

(N.1) $a_{jk} = \mathrm{I}\{|r_{jk}| > r\}$, where r is the cutoff calculated from the Fisher transformation;

(N.2) $a_{jk} = \mathrm{I}\{\pi_{jk} > \pi\}$, where π is the cutoff calculated from permutation which corresponds to the null that all covariates are not associated with response;

(N.3) $a_{jk} = \frac{1}{1+e^{-\alpha(\pi_{jk}-\pi)}}$, where $\alpha > 0$ can be determined by the scale-free topology criterion and π is defined in N.2;

(N.4) $a_{jk} = \pi_{jk}^{\alpha}$, where α is defined in N.3;

(N.5) $a_{jk} = \pi_{jk}^{\alpha} \mathrm{I}\{\pi_{jk} > \pi\}$, with α and π defined in N.3 and N.2, respectively;

(N.6) $a_{jk} = |r_{jk}| \mathrm{I}\{|r_{jk}| > r\}$ with r defined in N.1;

(N.7) $a_{jk} = \pi_{jk} \mathrm{I}\{\pi_{jk} > \pi\}$ with π defined in N.2.

... and many more possibilities!

Contrasted penalized estimation

$$
\begin{aligned}
P_{\lambda,\gamma}(\boldsymbol{\beta}) &= \sum_{j=1}^{p} \rho\left(||\boldsymbol{\beta}_j||_{1(2)}; \sqrt{M_j}\lambda_1, \gamma\right) \\
&+ \frac{1}{2}\lambda_2 d \sum_{1\leq j<k\leq p} a_{jk} \left(\frac{||\boldsymbol{\beta}_j||}{\sqrt{M_j}} - \frac{||\boldsymbol{\beta}_k||}{\sqrt{M_k}}\right)^2 .
\end{aligned}
$$

Rationale: penalize the contrast between $\boldsymbol{\beta}_j$ and $\boldsymbol{\beta}_k$; smooth over adjacent covariates.

A more familiar formulation

Denote $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' = \left(\frac{||\boldsymbol{\beta}_1||}{\sqrt{M_1}}, \cdots, \frac{||\boldsymbol{\beta}_p||}{\sqrt{M_p}}\right)'$. We express the the second penalty term using a positive semi-definite matrix L, which satisfies

$$\boldsymbol{\theta}' L \boldsymbol{\theta} = \sum_{1 \le j < k \le p} a_{jk} \left(\theta_j - \theta_k\right)^2, \forall \boldsymbol{\theta} \in \mathbf{R}^p .$$

Let $A = (a_{jk}, 1 \le j, k \le p)$ and $G = \text{diag}(g_1, \ldots, g_p)$, where $g_j = \sum_{k=1}^{p} a_{jk}$.

In a network where a_{jk} is the weight of edge (j, k), g_j is the degree of vertex j. We then have $\sum_{1 \le j < k \le p} a_{jk}(\theta_j - \theta_k)^2 = \boldsymbol{\theta}'(G - A)\boldsymbol{\theta}$. Thus, $L = G - A$.

$$P_{\lambda,\gamma}(\boldsymbol{\theta}) = \sum_{j=1}^{p} \rho(\theta_j; \lambda_1, \gamma) + \frac{1}{2}\lambda_2 d \boldsymbol{\theta}' L \boldsymbol{\theta}.$$

Computational algorithm

Computational algorithm With a sparse group Lasso penalty in the analysis of a single dataset under the linear regression model, an efficient algorithm based on coordinate descent is presented in Friedman et al. (2010). Unfortunately, because of the significant differences in penalties, that algorithm is not directly applicable under the present setup.

We first consider the integrative analysis of multiple studies with continuous responses and linear regression models. Using notations similar to those in Sectio

$Y = X\beta + \epsilon$. Here the objective function is

$$R(\beta) = \frac{1}{2n}||Y - \sum_{j=1}^{d} X[\cdot, j]\beta_j||_2^2 + \sum_{j=1}^{d}\left\{\rho(||\beta_j||_{\Sigma_j}; \sqrt{d_j}\lambda_1, a)\right.$$

In integrative analysis, multiple studies are independent. In add tion, we can carry out within-study standardization so that all co variance one. Under such conditions, Σ_js are diagonal, $j = 1, \ldots$ algorithm, consider update of estimate for a single group.

Given the group parameter vectors β_k $(k \neq j)$ fixed at their cur minimize the objective function $R(\beta)$ with respect to the jth grou objective function

$$R(\beta_j) = \frac{1}{2n}\|r_{-j} - X[\cdot, j]\beta_j\|_2^2 + \rho(\|\beta_j\|_{\Sigma_j}; \sqrt{d_j}\lambda_1, a) +$$

where $r_{-j} = Y - \sum_{k \neq j} X[\cdot, k]\tilde{\beta}_k$. The first order derivative of expr

$$\frac{\partial R(\beta_j)}{\partial \beta_j} = -\frac{1}{n}X[\cdot, j]'r_{-j} + \frac{1}{n}X[\cdot, j]'X[\cdot, j]\beta_j + \frac{\Sigma_j\beta_j}{\|\beta_j\|_{\Sigma_j}}\begin{cases}\sqrt{d_j}\lambda_1 - \frac{\|\beta_j\|_2}{a}, & \text{if } \|\beta_j\|_2 \leq \\ 0, & \text{if } \|\beta_j\|_2 > \end{cases}$$

where

$$t = \left(sgn(\beta_j^1)\begin{cases}\lambda_2 - \frac{|\beta_j^1|}{a}, & \text{if } |\beta_j^1| \leq b\lambda_2 \\ 0, & \text{if } |\beta_j^1| > b\lambda_2\end{cases}, \ldots, sgn(\beta_j^M)\begin{cases}\lambda_2 - \frac{|\beta_j^M|}{a}, & \text{if } |\beta_j^M| \leq b\lambda_2 \\ 0, & \text{if } |\beta_j^M| > b\lambda_2\end{cases}\right)'.$$

With standardization, $n^{-1}X[\cdot, j]'X[\cdot, j] = diag(n^1, \ldots, n^M)$. By setting expression (6) be zero, we have:

$$-z_j + g\beta_j + t = 0, \tag{7}$$

where $z_j = n^{-1}X[\cdot, j]'Y$, $g = diag(g_1, \ldots, g_M) = \Sigma_j h$ and

$$h = \left(1 + \frac{1}{\|\beta_j\|_{\Sigma_j}}\right)\begin{cases}\sqrt{d_j}\lambda_1 - \frac{\|\beta_j\|_{\Sigma_j}}{a}, & \text{if } \|\beta_j\|_{\Sigma_j} \leq a\sqrt{d_j}\lambda_1 \\ 0, & \text{if } \|\beta_j\|_{\Sigma_j} > a\sqrt{d_j}\lambda_1\end{cases}.$$

Denote z_j^k as the kth element of z_j. In g, first fix β_j at the current estimate $\tilde{\beta}_j$. in equation (7) can be rewritten as:

$$-\frac{z_j^k}{g_k} + \beta_j^k + sgn(\beta_j^k)\begin{cases}\frac{\lambda_2}{g_k} - \frac{|\beta_j^k|}{bg_k}, & \text{if } |\beta_j^k| \leq b\lambda_2 \\ 0, & \text{if } |\beta_j^k| > b\lambda_2\end{cases} = 0.$$

The solution to equation (8) is

$$\widehat{g_k\beta_j^k} = \begin{cases}\frac{S_1(z_j^k, \lambda_2)}{1 - \frac{1}{bg_k}}, & \text{if } |z_j^k| \leq b\lambda_2 g \\ z_j^k, & \text{if } |z_j^k| > b\lambda_2 g.\end{cases}$$

Here $S_1(z, \lambda) = sgn(z)(|z| - \lambda)_+$. For $k = 1, \ldots, M$, set $s_k = \widehat{g_k\beta_j^k}$ and $s = (s_1, .$ s back into its definition,

$$\Sigma_j^{\frac{1}{2}}\beta_j + \frac{\Sigma_j^{\frac{1}{2}}\beta_j}{\|\beta_j\|_{\Sigma_j}}\begin{cases}\sqrt{d_j}\lambda_1 - \frac{\|\beta_j\|_{\Sigma_j}}{a}, & \text{if } \|\beta_j\|_{\Sigma_j} \leq a\sqrt{d_j}\lambda_1 \\ 0, & \text{if } \|\beta_j\|_{\Sigma_j} > a\sqrt{d_j}\lambda_1\end{cases} = \Sigma_j^{-\frac{1}{2}}s.$$

Let $\theta_j = \Sigma_j^{\frac{1}{2}}\beta_j$. Expression (9) can be solved in a similar manner as with the gMCP, leading to

$$\hat{\theta}_j = \begin{cases}\frac{a}{a-1}S_2(\Sigma_j^{-\frac{1}{2}}s, \sqrt{d_j}\lambda_1), & \text{if } \|\Sigma_j^{-\frac{1}{2}}s\|_2 \leq a\sqrt{d_j}\lambda_1 \\ \Sigma_j^{-\frac{1}{2}}s, & \text{if } \|\Sigma_j^{-\frac{1}{2}}s\|_2 > a\sqrt{d_j}\lambda_1\end{cases},$$

where $S_2(z, t) = \left(1 - \frac{t}{\|z\|_2}\right)_+ z$ and

$$\hat{\beta}_j = \Sigma_j^{-\frac{1}{2}}\hat{\theta}_j. \tag{10}$$

Solving equation (7) amounts to iteratively calculating equations (8) to (10) until convergence. Overall, consider the following algorithm. With fixed tuning parameters,

1. Initialize $s = 0$, the estimate $\beta^{(0)} = (0, \ldots, 0)'$, and the vector of residuals $r = Y - X\beta^{(0)}$;
2. For $j = 1, \ldots, d$,
   (a) Compute $\beta_j^{(s+1)}$. This is achieved by solving equation (7), which amounts to iterating equations (8) to (10) until convergence. In our numerical studies, convergence is achieved for all datasets within ten iterations;
   (b) Update $r \leftarrow r - X[\cdot, j](\beta_j^{(s+1)} - \beta_j^{(s)})$;

   Update $s \leftarrow s + 1$;
3. Repeat Step 2 until convergence.

Computation is realized using an iterative coordinate descent algorithm.

With coordinate descent, we update the estimate for one group of coefficients at a time, and cycle through all groups. This process is repeated until convergence.

$R(\boldsymbol{\beta})$ and the second penalty have local quadratic forms. Their sum (f) is regular in the sense of Tseng (2001).

The coordinate descent solution converges to a coordinate-wise minimum point of f, which is also a stationary point.

Simulation

Three datasets; 100 samples per dataset; 500 covariates per subject.

500 covariates belong to 100 clusters. Covariates within the same clusters are correlated. Different clusters are independent.

Among the 500 covariates, 20 (4 clusters) have nonzero regression coefficients.

Nonzero regression coefficients $\sim Unif[0.25, 0.75]$. Random error $\sim N(0, 1)$. Log censoring time: normally distributed.

benchmark	N.1	N.2	N.3	N.4	N.5	N.6	N.7
			$\rho = 0.1$				
19.5	19.5	19.3	19.7	19.4	19.4	19.4	19.5
11.8	13.3	13.3	16.8	16.2	12.9	13.5	14.1
4.5	4.6	4.6	4.4	4.5	4.5	4.7	4.6
			$\rho = 0.5$				
18.9	19.9	20.0	19.1	19.9	19.9	20.0	20.0
11.3	18.4	11.0	16.0	16.7	9.9	17.8	12.3
4.5	3.7	3.7	4.1	3.7	3.8	3.7	3.7
			$\rho = 0.9$				
11.1	19.6	20.0	19.3	19.9	19.8	20.0	20.0
7.5	7.8	3.5	21.1	5.1	2.5	7.1	2.5
4.9	3.9	3.8	4.2	3.7	3.9	3.7	3.8

The first row is number of true positives, the second row is number of false positives, and the third row is mean prediction error.

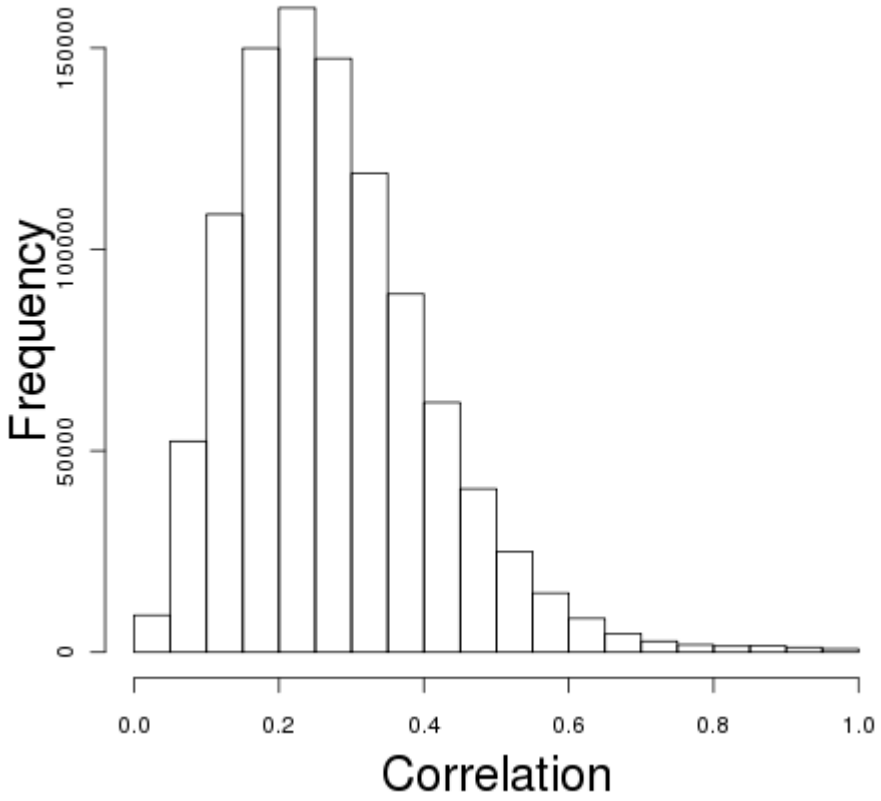
Analysis of lung cancer prognosis studies

Lung cancer is the leading cause of death from cancer for both men and women in the US. NSCLC accounts for up to 85% of lung cancer deaths.

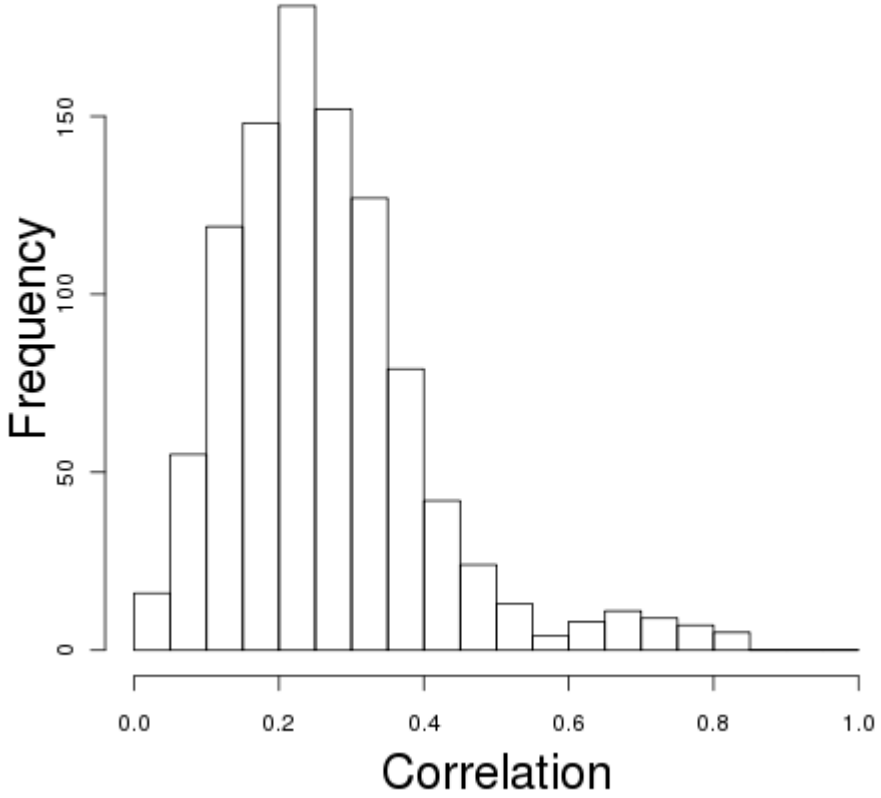
The UM (University of Michigan) study has 175 patients and 102 deaths. Median follow-up=53 months. The HLM (Moffitt Cancer Center) study has 79 subjects and 60 deaths. Median follow-up=39 months. The CAN/DF (Dana-Farber) study has 82 patients and 35 deaths. Median follow-up=51 months.

22,283 probe sets were profiled. We rank the probe sets using their variations and select the top 1,000 probes for downstream analysis.

Canonical correlation among all genes



Canonical correlation with probe 206561_s_at



Numbers of identified genes and overlaps

	N.1	N.2	N.3	N.4	N.5	N.6	N.7	gMCP
N.1	22	19	22	13	14	13	13	9
N.2		20	19	14	15	14	14	8
N.3			25	14	15	14	14	9
N.4				15	15	14	14	5
N.5					16	15	15	5
N.6						19	19	4
N.7							20	4
gMCP								9

Evaluation of prediction performance

We generate training sets and testing sets by random splitting with sizes 2:1. Estimates are generated using the training sets only. We then make prediction for subjects in the testing sets. With the predicted linear risk scores $\mathbf{X}\hat{\boldsymbol{\beta}}$, dichotomize at the median, create two risk groups, and compute the logrank statistic, which measures the difference in survival between the two groups.

The average logrank statistics over 100 splits are 4.47 (N.1), 4.30 (N.2), 4.77 (N.3), 4.93 (N.4), 4.23 (N.5), 5.13 (N.6) and 4.03 (N.7) for SGLS and 3.77 for gMCP.

Genes identified by SGLS (N.6) but not gMCP

Gene SCGB1A1 (secretoglobin, family 1A, member 1), GPX2 (glutathione peroxidase 2), ABP1 (amiloride binding protein 1), CST1 (cystatin SN), TSPYL5 (testis-specific Y-encoded-like protein 5), ID1 (inhibitor of DNA binding 1, dominant negative helix-loop-helix protein), TUBB2A (tubulin, beta 2A class IIa), GEM (GTP binding protein overexpressed in skeletal muscle), KAL1 (Kallmann syndrome 1 sequence), PAH (phenylalanine hydroxylase), LYZ (lysozyme), PNMAL1(paraneoplastic Ma antigen family-like), ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) and C4BPB (complement component 4 binding protein, beta).

Searching published literature suggests that these genes may have important implications.

Across-datasets structure

Here the structure describes the relationships among regression coefficients for the same covariate across multiple datasets.

Consider a simulation scenario where the regression coefficients of response-associated covariates are the same across multiple datasets.

True	Benchmark			Contrasted penalization		
	D1	D2	D3	D1	D2	D3
0.4	0.186	0.391	0.112	0.302	0.292	0.317
0.5	0.349	0.400	0.465	0.411	0.428	0.537
0.6	0.587	0.244	0.392	0.553	0.461	0.587
0.7	0.592	0.746	0.553	0.637	0.659	0.695
0.8	0.683	0.769	0.698	0.617	0.661	0.732
-0.4	-0.302	-0.312	-0.187	-0.309	-0.253	-0.287
-0.5	-0.627	-0.519	-0.482	-0.599	-0.575	-0.502
-0.6	-0.558	-0.742	-0.514	-0.583	-0.568	-0.599
-0.7	-0.571	-0.576	-0.556	-0.557	-0.612	-0.600
-0.8	-0.635	-0.622	-0.495	-0.704	-0.624	-0.730

Under certain scenarios (for example when multiple datasets are independently generated under the same protocol), it is reasonable to expect "similar" regression coefficients across multiple datasets.

However, in practice, we never know how "close" two datasets are. We cannot rule out the scenario where a covariate has effects with conflicting signs.

Penalized estimation

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho\left(||\boldsymbol{\beta}_j||_{1(2)}; \sqrt{M_j}\lambda_1, \gamma\right) + \lambda_2 \sum_{j=1}^{d} \sum_{(k,l):k\neq l} a_j^{kl}(\beta_j^k - \beta_j^l)^2.$$

$$a_j^{kl} = I\{\text{sgn}(\beta_j^k) = \text{sgn}(\beta_j^l)\}$$

where sgn is the sign function. $\lambda_2 \geq 0$ is a data-dependent tuning parameter.

When $\text{sgn}(\beta_j^k) \neq \text{sgn}(\beta_j^l)$, covariate j demonstrates different effects in different studies. In this case, the contrast has no effect.

When $\text{sgn}(\beta_j^k) = \text{sgn}(\beta_j^l)$, covariate j has qualitatively similar effects in studies k and l. The contrast penalty shrinks the difference between β_j^k and β_j^l and encourages them to be similar.

The "smoothing" structure is mainly for covariates with nonzero effects. We may further consider

$$a_j^{kl} = I(||\beta_j||_2 \neq 0) \times I\{\text{sgn}(\beta_j^k) = \text{sgn}(\beta_j^l)\}.$$

With practical data, sgn(β_j^k) needs to be estimated. There are several proposals (a) marginal estimation (in the spirit of screening), (b) single-dataset penalization, (c) integrative penalization, etc.

Their asymptotic consistency can be established.

Our limited experience suggests that (b) and (c) work reasonably well.

Computation is also based on coordinate descent, in a similar manner as with the previous estimate.

Simulation study

Three datasets are simulated, each with 100 subjects. For each subject, $d = 1,000$ covariates are simulated to have a multivariate normal distribution.

Under the heterogeneity model, all three datasets share five common markers. In addition, each dataset has five dataset-specific markers. Thus, across the three datasets, there are a total of 30 markers.

In addition, as a special case of the heterogeneity model, we also consider the homogeneity model.

The regression coefficients of the response-associated covariates are

(0.4, 0.5, 0.6, 0.7, 0.8, -0.4, -0.5, -0.6, -0.7, -0.8),

(0.4, -0.5, 0.6, -0.7, 0.8, -0.4, 0.5, -0.6, 0.7, -0.8),

(0.4, 0.5, 0.6, 0.7, 0.8, -0.4, -0.5, -0.6, -0.7, -0.8)

for dataset 1-3, respectively.

Heterogeneity model

	Contrasted ($\lambda_2 =$)				
Benchmark	0.01	0.1	1	10	100
Auto-regressive $\rho = 0.2$					
10.5	10.4	10.7	10.8	10.8	10.6
46.4	45.5	42.8	41.2	42.0	39.2
7.6	7.0	6.3	6.0	6.0	6.0
Auto-regressive $\rho = 0.8$					
7.4	7.5	7.8	7.9	7.8	7.7
34.5	28.1	27.2	27.9	26.8	24.5
6.7	6.6	6.0	6.3	6.7	7.3
Banded scenario 1					
8.5	8.5	8.5	8.8	8.6	8.7
45.2	41.2	40.7	41.4	39.1	37.9
7.4	7.1	6.4	6.3	6.2	6.4
Banded scenario 2					
8.9	9.1	9.5	9.0	8.8	8.8
46.3	41.6	42.2	40.2	38.5	37.6
7.4	7.0	6.6	6.4	6.4	6.7

True positive; False positive; Prediction SSE.

Homogeneity model

	Contrasted ($\lambda_2 =$)				
Benchmark	0.01	0.1	1	10	100
Auto-regressive $\rho = 0.2$					
25.0	25.1	25.1	25.2	25.1	25.1
32.9	25.0	25.9	22.2	21.1	18.9
2.6	2.4	2.1	2.2	2.2	2.2
Auto-regressive $\rho = 0.8$					
14.4	14.9	15.1	15.3	14.9	14.5
24.0	20.0	15.7	12.9	11.4	9.9
2.5	2.3	2.3	3.1	3.4	3.5
Banded scenario 1					
24.7	24.5	24.4	24.2	24.2	24.1
37.4	29.8	27.9	20.4	18.8	17.3
2.7	2.5	2.3	2.4	2.4	2.5
Banded scenario 2					
21.9	22.0	21.9	21.7	21.6	21.2
25.9	19.2	15.2	14.2	13.0	11.7
2.2	2.0	2.0	2.4	2.7	2.8

A generic framework

$$\widehat{\boldsymbol{\beta}} = argmin\,\{R(\boldsymbol{\beta}) + P(\boldsymbol{\beta}) + P_c(C\boldsymbol{\beta})\}$$

The matrix C specifies the contrasts. It describes the network structure among all covariates (in the same or different datasets).

Consider $\boldsymbol{\beta}^o$, a $p \times M$ matrix. Its components that correspond to the zero components of $\boldsymbol{\beta}$ have been set as zero. That is, selection has been pre-conducted by an "oracle". Consider

$$\widehat{\boldsymbol{\beta}}^o = argmin\,\{R(\boldsymbol{\beta}) + P_c(C\boldsymbol{\beta})\}$$

Assume C is known.

Some Assumptions

M is fixed. $\log(p)/\sum_m n^m \to 0$.

The size of set $\{j : ||\beta_j||_2 > 0\}$ is finite.

All X^ms satisfy the SRC (sparse Rietz condition): all submatrices with sizes smaller than a fixed value have bounded eigenvalues.

Model specific assumptions: for example if $Y^m = \alpha^m + \beta^{m\prime} X^m + \epsilon^m$, then ϵ^m has a sub-Gaussian distribution.

Main result: $Pr(\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^o) \to 1$

The unsolved, real problem

The contrast penalty C can be data-dependent.

Consider for example network-based analysis. C is closely related to the variance-covariance matrix. If no additional assumption is made and $p >> n$, we may not have consistent estimates (of, for example, eigenvalues and eigenvectors).

We can fix the above inconsistency problem by imposing, for example, the banded structure. However, we do not know what $\widehat{\boldsymbol{\beta}}$ will be like with estimated C.

Remarks

In high-dimensional data analysis, contrasted penalization provides an effective way to accommodate secondary data structures.

When the contrast is properly specified, simulation shows that contrasted penalization can improve selection and estimation. However, it is non-trivial to specify C.

Need development in these aspects: conceptual, theoretical, and computational.

Acknowledgements

Collaborators: Dr. Jin Liu (UIC), Mr. Xingjie Shi (Yale), Prof. Jian Huang (UIowa).

Funding support from NIH and Yale University.

Many thanks to the workshop organizers.