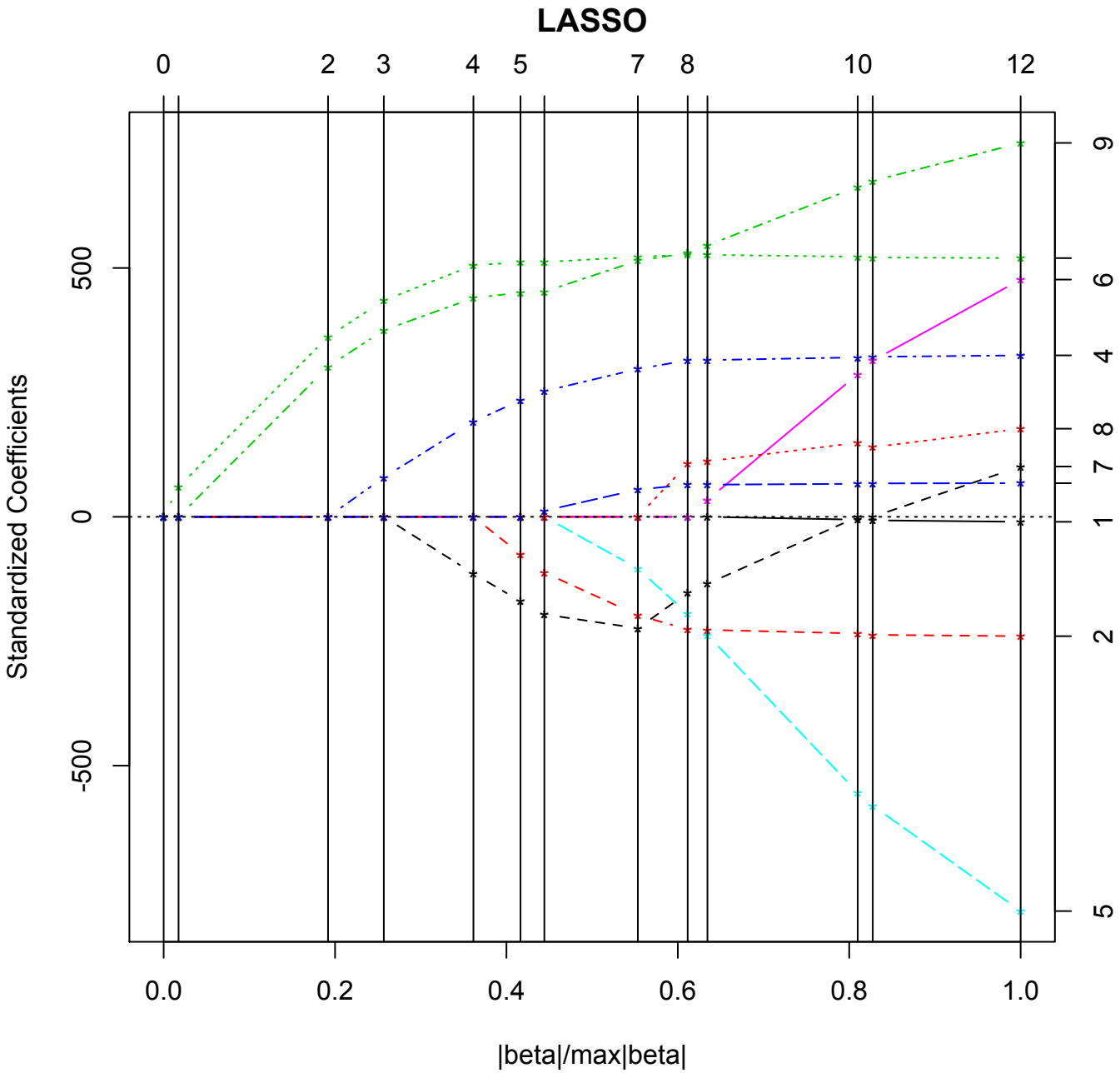
Interplay between Statistics and Optimization

Hui Zou

University of Minnesota

SAMSI

August 29, 2016


LASSO
0
2
3
4
5
7
8
10
12
Standardized Coefficients
500
0
-500
9
6
4
8
7
1
2
5
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|

- Microarray data in early 2000s
  - Large scale multiple testing–false discovery rate (FDR) control (Benjamini and Hochberg), local FDR (Efron), Higher criticisim (Donoho and Jin), SAM (Tibshirani)
  - Regression analysis–Least Angle Regression (Efron, Hastie, Johnstone and Tibshirani), Lasso
- Various penalization techniques: SCAD, MCP, Elastic Net, Adaptive Lasso, fused Lasso, group Lasso,...
- More sophisticated models/problems: GLM, GAM, precision matrix estimation, covariance matrix estimation, ...
- Compressed sensing, Matrix completion, Robust PCA
- Tensor regression, Tensor completion, Tensor decomposition

My personal view

- Optimization for Statistics: model fitting, model formulation, theoretical analysis
- Statistics for Optimization: new research thrusts

Today's talk:

- Majorization-Minimization (MM)
- Alternating Direction Method of Multipliers (ADMM)

Majorization-Minimization

Solve $\operatorname{argmin}_{\theta} C(\theta)$

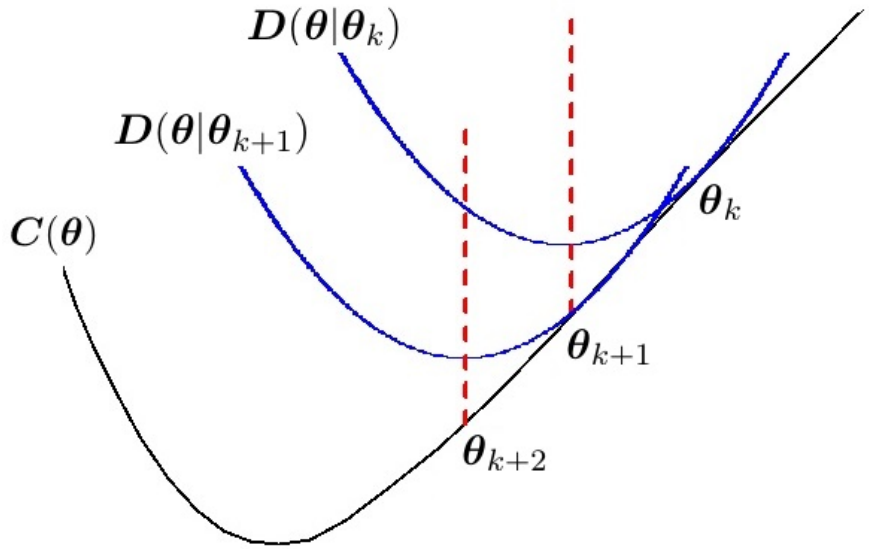
Majorization step:

$C(\theta) < D(\theta|\theta_k)$ for any $\theta \neq \theta_k$,

$C(\theta_k) = D(\theta_k|\theta_k)$

Minimization step:

$$\theta_k \leftarrow \theta_{k+1} = \operatorname*{argmin}_{\theta} D(\theta|\theta_k)$$



Lange, Hunter and Yang (2000)[optimization transfer]; Hunter & Lange (2000) [MM]; Wu and Lange (2010) [EM and MM]

LLA

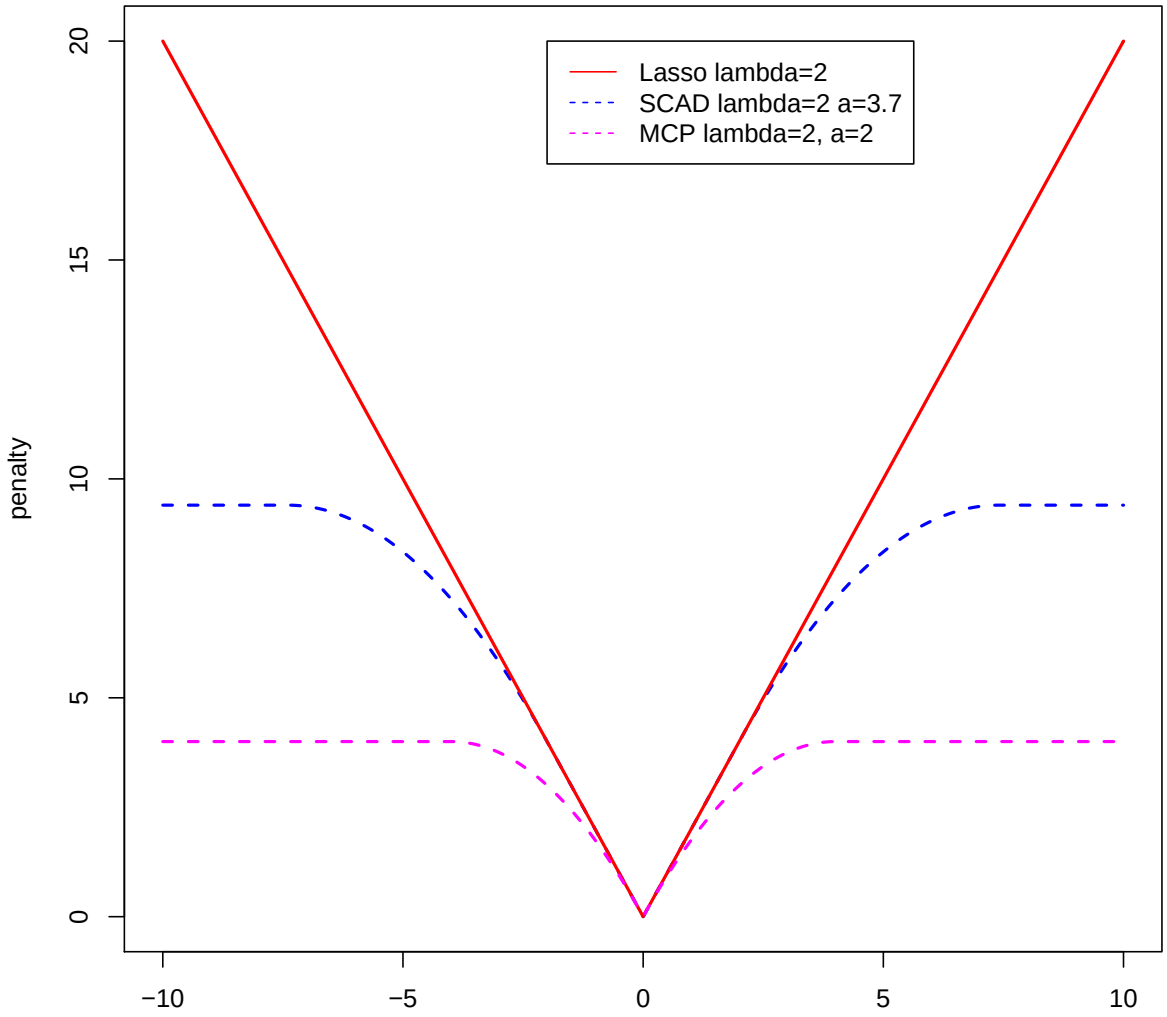
Zou and Li (2008)

Fan, Xue and Zou (2014)

Nonconvex penalized regression

$$\min_{\boldsymbol{\beta}} \ \ell_n(\boldsymbol{\beta}) + \sum_j P_\lambda(|\beta_j|)$$

- ℓ_n is convex and represents the statistical inference model
  - least squares loss
  - Huber 's M loss or least absolute loss
  - logistic regression: negative log-Bernoulli-likelihood
  - quantile regression: check loss
  - Ising model: composite conditional likelihood (Xue, Zou and Cai, 2012)
- $P_\lambda(t)$ is a non-decreasing concave function for $t \in (0, \infty)$
  - L_q norm penalty $(0 < q < 1)$
  - SCAD (Fan and Li, 2001)
  - MCP (Zhang, 2010)


Lasso lambda=2
SCAD lambda=2 a=3.7
MCP lambda=2, a=2
penalty
0
5
10
15
20
−10
−5
0
5
10

LLA

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_\lambda(\beta_j) \right\}$$

1. Start with some initial estimator $\boldsymbol{\beta}^{(0)}$.
2. At step k, define

$$Q_\lambda(\beta_j) = P_\lambda(|\beta_j^{(k)}|) + P'_\lambda(|\beta_j^{(k)}|+)(|\beta_j| - |\beta_j^{(k)}|)$$

3. Solve $\boldsymbol{\beta}^{(k+1)} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{d} Q_\lambda(\beta_j) \right\}$.

Iterate between Steps 2 and 3.

LLA and EM

Condition on P_λ: if there is a positive function $H(t)$ such that

$$\exp(-nP_\lambda(|\beta|)) = \int_0^\infty H(t)e^{-t|\beta|}dt. \quad (*)$$

Let $\pi(t) = \frac{2}{t}H(\frac{1}{t})$ and $p(\beta_j|\tau_j) = \frac{1}{2\tau_j}e^{-\frac{|\beta_j|}{\tau_j}}$. Then $(*)$ yields

$$\exp(-nP_\lambda(|\beta_j|)) = \int_0^\infty p(\beta_j|\tau_j)\pi(\tau_j)d\tau_j. \quad (**)$$

$(**)$ represents a hierarchical Bayesian model and suggests an EM algorithm for maximizing the penalized likelihood by treating τ_js as “missing values".

Under condition $(*)$ EM=LLA.

The issue of multiple local solutions

- The folded concave penalization problem usually has multiple local solutions, but the theory (namely, the oracle property) is established only for one of the unknown local solutions (Fan and Li, 2001; Fan and Peng, 2004; Lv and Fan, 2008; Fan and Lv, 2011; ...).

- Over a decade, the challenging fundamental issue still remains that it is not clear whether the local optimal solution computed by a given optimization algorithm possesses those nice theoretical properties.

Numeric demonstration

Simulation model: $y \sim \text{Bernoulli}(\frac{\exp(X\beta^\star)}{1+\exp(X\beta^\star)})$, where $X \sim N_p(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$ and $\beta^\star = (3, 1.5, 0, 0, 2, 0_{p-5})$.

	$n = 200$ & $p = 1000$			
	ℓ_1 loss	ℓ_2 loss	# FP	# FN
	Sparse logistic regression			
Lasso	5.67	2.37	24.02	0.04
	(0.05)	(0.02)	(0.44)	(0.01)
SCAD-CD	4.50	2.13	13.99	0.08
	(0.06)	(0.02)	(0.31)	(0.01)
SCAD-LLA-zero	2.16	1.32	0.31	0.22
	(0.11)	(0.06)	(0.05)	(0.02)
SCAD-LLA-Lasso	2.08	1.28	0.26	0.19
	(0.10)	(0.06)	(0.04)	(0.02)

LLA closes the theoretical gap

In Fan, Xue and Zou (2014) it is shown that

Theorem

- ⋆ *If the initial estimator is Lasso, then the two-step LLA procedure finds the oracle solution with high probability.*
- ⋆ *If the initial estimator is zero, then the three-step LLA procedure finds the oracle solution with high probability.*

As illustration, the theory is verified for penalized least squares, penalized logistic regression, penalized quantile regression and penalized graphical model estimation.

The philosophical root of our theory

- In the classical MLE theory, when the log-likelihood function is not concave, one of the local maximizers of the log-likelihood function is shown to be asymptotic efficient, but how to compute that estimator is very challenging and often unclear.
- Le Cam (1956) (and later Bickel 1975) overcame this technical difficulty by focusing on a specially designed one-step Newton-Raphson estimator initialized by a root-n estimator.
- Le Cam did not try to get the global maximizer nor the theoretical local maximizer of the likelihood.

The search for the global minimizer

Mixed Integer Programming has been used to get the global minimizer of L_0 penalized and SCAD penalized least squares.

- Dimitris Bertsimas, Angela King and Rahul Mazumder (2016, AoS). Best Subset Selection via a Modern Optimization Lens.
- Hongcheng Liu, Tao Yao, Runze Li (2016). Global solutions to folded concave penalized nonconvex learning. AoS, 44(2), 629-659.

Extension to more general models?

BMD

Yang and Zou (2013)

Coordinate descent for lasso

$$\underset{\beta_1,\ldots,\beta_p}{\operatorname{argmin}} f(\beta_1,\ldots,\beta_p) + \sum_{j=1}^{p} \lambda|\beta_j|$$

1. Initialization of $\tilde{\beta}$
2. Cyclic coordinate descent: for $j = 1, 2, \ldots, p, 1, 2, \ldots$, update β_j by minimizing the objective function

$$\tilde{\beta}_j^{update} \leftarrow \underset{\beta_j}{\operatorname{argmin}} f(\tilde{\beta}_1,\ldots,\tilde{\beta}_{j-1},\beta_j,\tilde{\beta}_{j+1},\tilde{\beta}_p) + \lambda|\beta_j|$$

3. Repeat (2) till convergence.

Lasso regression: $f(\beta_1, \ldots, \beta_p) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$

$$\tilde{\beta}_j^{update} \leftarrow \underset{\beta_j}{\operatorname{argmin}} \|\mathbf{y} - \sum_{k \neq j} x_k \tilde{\beta}_k - x_j \beta_j\|_2^2 + \lambda |\beta_j|$$

reduces to soft-thresholding.

- Fu (1998) proposed the algorithm named “shooting".
- Friedman, Hastie and Tibshirani (2008) glmnet, the same CD but with clever implementation tricks such as active set, warm start and later strong rule.
- For lasso logistic regression, Friedman, Hastie and Tibshirani (2008) did CD within a Newton-Ralphson loop. Genkin, Lewis and Madigan (2007) did the standard CD by solving the one-dimensional optimization repeatedly.

Group lasso regression

$$\min_{(\beta_0,\boldsymbol{\beta})} \left[\frac{1}{2} \left\| \mathbf{y} - \beta_0 - \sum_k \mathbf{X}_{(\mathbf{k})} \boldsymbol{\beta}^{(k)} \right\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{p_k} \| \boldsymbol{\beta}^{(k)} \|_2 \right] .$$

- Group Lasso penalty was introduced in Turlach, Vanebles and Wright (2004) and Yuan and Lin (2006).
- A blockwise descent algorithm under a groupwise orthonormal condition: $\mathbf{X}_{(\mathbf{k})}$ columns are orthonormal.
- Orthonormal condition is incompatible with cross-validation, bootstrap, sub-sampling.

A general group lasso problem

$$\arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \tau_i \Phi(y_i, \beta^{\mathsf{T}} \mathbf{x}_i) + \lambda \sum_{k=1}^{K} w_k \|\beta^{(k)}\|_2$$

where $\tau_i \geq 0$ and $w_k \geq 0$ for all i, k.

- The observation weights τ_is are introduced in order to cover methods such as weighted regression and weighted large margin classification (biased sampling, unequal cost classification).
- The penalty weights w_ks make a more flexible model. The default choice for w_k is $\sqrt{p_k}$. If we do not want to penalize a group of predictors, simply let the corresponding weight be zero.

Loss functions

- Least squares: $\Phi(y,f) = \frac{1}{2}(y-f)^2$
- Logistic regression: $\Phi(y,f) = \log(1+e^{-yf}), y = \pm 1$
- Squared hinge loss: $\Phi(y,f) = [(1-yf)_+]^2, y = \pm 1$
- Huberized SVM loss: $\Phi(y,f) = \mathrm{hsvm}(yf), y = \pm 1$ where

$$\mathrm{hsvm}(t) = \begin{cases} 0, & t > 1 \\ (1-t)^2/2\delta, & 1-\delta < t \leq 1 \\ 1-t-\delta/2, & t \leq 1-\delta. \end{cases}$$

Let $\mathbf{D}$ denote the data $\{\mathbf{y}, \mathbf{X}\}$ and define

$$L(\boldsymbol{\beta} \mid \mathbf{D}) = \frac{1}{n} \sum_{i=1}^{n} \tau_i \Phi(y_i, \boldsymbol{\beta}^\mathsf{T} \mathbf{x}_i).$$

Definition

The loss function Φ is said to satisfy the QM condition, if

(i). $\nabla L(\boldsymbol{\beta} | \mathbf{D})$ exists everywhere.

(ii). There exists a $p \times p$ matrix $\mathbf{H}$, which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$,

$$\begin{aligned} L(\boldsymbol{\beta} \mid \mathbf{D}) \quad \leq \quad & L(\boldsymbol{\beta}^* \mid \mathbf{D}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\mathsf{T} \nabla L(\boldsymbol{\beta}^* | \mathbf{D}) \\ & + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\mathsf{T} \mathbf{H} (\boldsymbol{\beta} - \boldsymbol{\beta}^*). \end{aligned}$$

Loss	$-\nabla L(\boldsymbol{\beta} \mid \mathbf{D})$	$\mathbf{H}$
Least squares	$\frac{1}{n}\sum_{i=1}^{n} \tau_i(y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta})\mathbf{x}_i$	$\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X}/n$
Logistic regression	$\frac{1}{n}\sum_{i=1}^{n} \tau_i y_i \mathbf{x}_i \frac{1}{1+\exp(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})}$	$\frac{1}{4}\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X}/n$
Squared hinge loss	$\frac{1}{n}\sum_{i=1}^{n} 2\tau_i y_i \mathbf{x}_i (1 - y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})_+$	$4\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X}/n$
Huberized SVM loss	$\frac{1}{n}\sum_{i=1}^{n} \tau_i y_i \mathbf{x}_i \text{hsvm}'(y_i \mathbf{x}_i^\intercal \boldsymbol{\beta})$	$\frac{2}{\delta}\mathbf{X}^\intercal \boldsymbol{\Gamma} \mathbf{X}/n$

$$\boldsymbol{\Gamma} = diag(\tau_1, \ldots, \tau_n)$$

Write $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^{(k')} = \widetilde{\boldsymbol{\beta}}^{(k')}$ for $k' \neq k$.

Given $\boldsymbol{\beta}^{(k')} = \widetilde{\boldsymbol{\beta}}^{(k')}$ for $k' \neq k$, the optimal $\boldsymbol{\beta}^{(k)}$ is defined as

$$\underset{\boldsymbol{\beta}^{(k)}}{\operatorname{argmin}} L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda w_k \|\boldsymbol{\beta}^{(k)}\|_2.$$

By QM condition,

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) + (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^{\mathsf{T}} \nabla L(\widetilde{\boldsymbol{\beta}} | \mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}})^{\mathsf{T}} \mathbf{H} (\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}).$$

Write $U(\widetilde{\boldsymbol{\beta}}) = -\nabla L(\widetilde{\boldsymbol{\beta}} | \mathbf{D})$.

$$\begin{aligned} L(\boldsymbol{\beta} \mid \mathbf{D}) \quad \leq \quad & L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^{\mathsf{T}} U^{(k)} \\ & + \frac{1}{2}(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^{\mathsf{T}} \mathbf{H}^{(k)} (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}). \end{aligned}$$

Let η_k be the largest eigenvalue of $\mathbf{H}^{(k)}$. We set $\gamma_k = (1 + 10^{-4})\eta_k$

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^{\intercal} U^{(k)} + \frac{1}{2}\gamma_k \|(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})\|_2^2 \quad (*)$$

"=" holds if only if $\boldsymbol{\beta}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}$

The minimizer of the right hand side of $(*)$ is

$$\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+ .$$

The whole process drives the objective strictly downhill unless the optimal solution is reached (i.e., KKT conditions are satisfied).

BMD for group lasso

- For $k = 1, \ldots, K$, compute γ_k, the largest eigenvalue of $\mathbf{H}^{(k)}$
- $\gamma_k = (1 + 10^{-4})\gamma_k$ (for nontrival groups with size ≥ 2)
- Initialize $\widetilde{\boldsymbol{\beta}}$.
- Repeat the following cyclic blockwise updates until convergence:
  - ⋆ for $k = 1, \ldots, K$, do (1)–(3)
    - ⋆ (1) Compute $U(\widetilde{\boldsymbol{\beta}}) = -\nabla L(\widetilde{\boldsymbol{\beta}}|\mathbf{D})$.
    - ⋆ (2) Compute
      $$\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda w_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+ .$$
    - ⋆ (3) Set $\widetilde{\boldsymbol{\beta}}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$.

`gglasso` package: also uses active set, strong rule and warm start.

Competitors

- block coordinate gradient descent `grplasso`: Meier et al. (2008) for group-lasso logistic regression.
- **ISTA-BC** algorithm: Qin et al. (2010), an extension of the ISTA/FISTA (Beck & Teboulle 2009) based on variable step-lengths.
- **SLEP** implemented Nesterov's method: Liu et al. (2009)

Dataset	Type	n	q	p	Data Source
Autompg	R	392	7	31	(Quinlan 1993)
Bardet	R	120	200	1000	(Scheetz et al. 2006)
Cardiomypathy	R	30	6319	31595	(Segal et al. 2003)
Spectroscopy	R	103	100	500	(Sabo et al. 2008)
Breast	C	42	22283	111415	(Graham et al. 2010)
Colon	C	62	2000	10000	(Alon et al. 1999)
Prostate	C	102	6033	30165	(Singh et al. 2002)
Sonar	C	208	60	300	(Gorman et al. 1988)

Some real datasets. n is the number of instances. q is the number of original variables. p is the number of predictors after expansion. “R" means regression and “C" means classification.

Group-lasso GAM regression, timing performance				
Dataset	*Autompg*	*Bardet*	*Cardiomypathy*	*Spectroscopy*
SLEP	3.14	9.96	78.23	9.37
ISTA-BC	5.66	1.55	2.43	1.31
gglasso	2.51	0.77	2.48	0.76

All experiments were carried out on an Intel Xeon X5560 (Quad-core 2.8 GHz) processor.

Group-lasso GAM classification, timing performance				
Dataset	*Colon*	*Prostate*	*Sonar*	*Breast*
grplasso (Logit)	60.42	111.75	24.55	439.76
SLEP (Logit)	75.31	166.91	5.49	358.75
gglasso (Logit)	1.13	3.877	1.54	9.62
gglasso (HSVM)	1.15	3.53	0.66	9.15

All experiments were carried out on an Intel Xeon X5560 (Quad-core 2.8 GHz) processor.

A Small Trick

Yang and Zou (2012)

A counterintuitive phenomenon

Consider the glmnet for fitting elastic net penalized regression. W.L.O.G. assume $\sum_{i=1}^N x_{ij} = 0$, $\frac{1}{N}\sum_{i=1}^N x_{ij}^2 = 1$, for $j = 1, \ldots, p$.

$$R(\beta_0, \beta) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^\intercal \beta)^2 + P_{\lambda,\alpha}(\beta),$$

where $P_{\lambda,\alpha}(\beta)$ is the elastic net penalty

$$P_{\lambda,\alpha}(\beta) \quad = \lambda \sum_{j=1}^{p} p_\alpha(\beta_j) = \quad \lambda \sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right].$$

`glmnet` implements the standard CD algorithm in which we iteratively solve a univariate elastic net problem

$$\hat{\beta}_j = \arg\min_{\beta_j} R(\beta_j|\tilde{\beta}_0, \tilde{\beta}),$$

where

$$R(\beta_j|\tilde{\beta}_0, \tilde{\beta}) = \frac{1}{2}\left(\beta_j - \tilde{\beta}_j\right)^2 - \frac{1}{N}\sum_{i=1}^{N} r_i x_{ij}\left(\beta_j - \tilde{\beta}_j\right) + \lambda p_\alpha(\beta_j).$$

$$\hat{\beta}_j = \frac{S\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i + \tilde{\beta}_j, \lambda\alpha\right)}{1 + \lambda(1-\alpha)},$$

where $S(z,t) = (|z| - t)_+ \text{sgn}(z)$.

A tiny change to glmnet

We change the univariate update formula to

$$\hat{\beta}_j^{\mathrm{B}} = \frac{S\left(\frac{1}{N}\sum_{i=1}^{N} x_{ij} r_i + f \cdot \tilde{\beta}_j, \lambda\alpha\right)}{f \cdot 1 + \lambda(1-\alpha)} \quad (f \geq 1)$$

- Yang made a code error by using $f = 2$ in glmnet, but still got good/even better results.
- As long as $f \geq 1$ the iterative process converges to the desired solution.
- A bigger f means a smaller step size along each coordinate direction. For an orthogonal design, $f = 1$ is the best choice.

Simulation

FHT model:

We simulated data with N observations and p predictors where each pair of predictors X_j and $X_{j'}$ have the same population correlation ρ, with ρ ranges from zero to 0.95.

The response variable was generated by

$$Y = \sum_{j=1}^{p} X_j \beta_j + k \cdot N(0,1),$$

where $\beta_j = (-1)^j \exp(-(2j-1)/20)$ and k is set to make the signal-to-noise ratio equal 3.

We compared $f = 1$ (glmnet) and $f = 2$ (glmnet2).

	Correlation					
	0	0.1	0.2	0.5	0.8	0.95
	$\alpha = 1$					
	$N = 100, p = 5000$					
glmnet	0.2222	0.2339	0.2979	0.4606	0.7919	1.9016
glmnet2	0.2533	0.2519	0.2886	0.3758	0.5450	1.0735
	$\alpha = 0.5$					
	$N = 100, p = 5000$					
glmnet	0.2107	0.2189	0.2356	0.3669	0.7765	2.1528
glmnet2	0.2225	0.2285	0.2414	0.2861	0.4876	1.3335

A simple explanation

$$v_j = (0, \cdots, \frac{x_j^\top y}{fN}, \cdots, 0) \qquad \mathbf{u}_j = (u_{kj})_{p\times 1} = \begin{cases} -\frac{1}{f} & k = j \\ -\frac{\rho}{f} & k \neq j \end{cases}$$

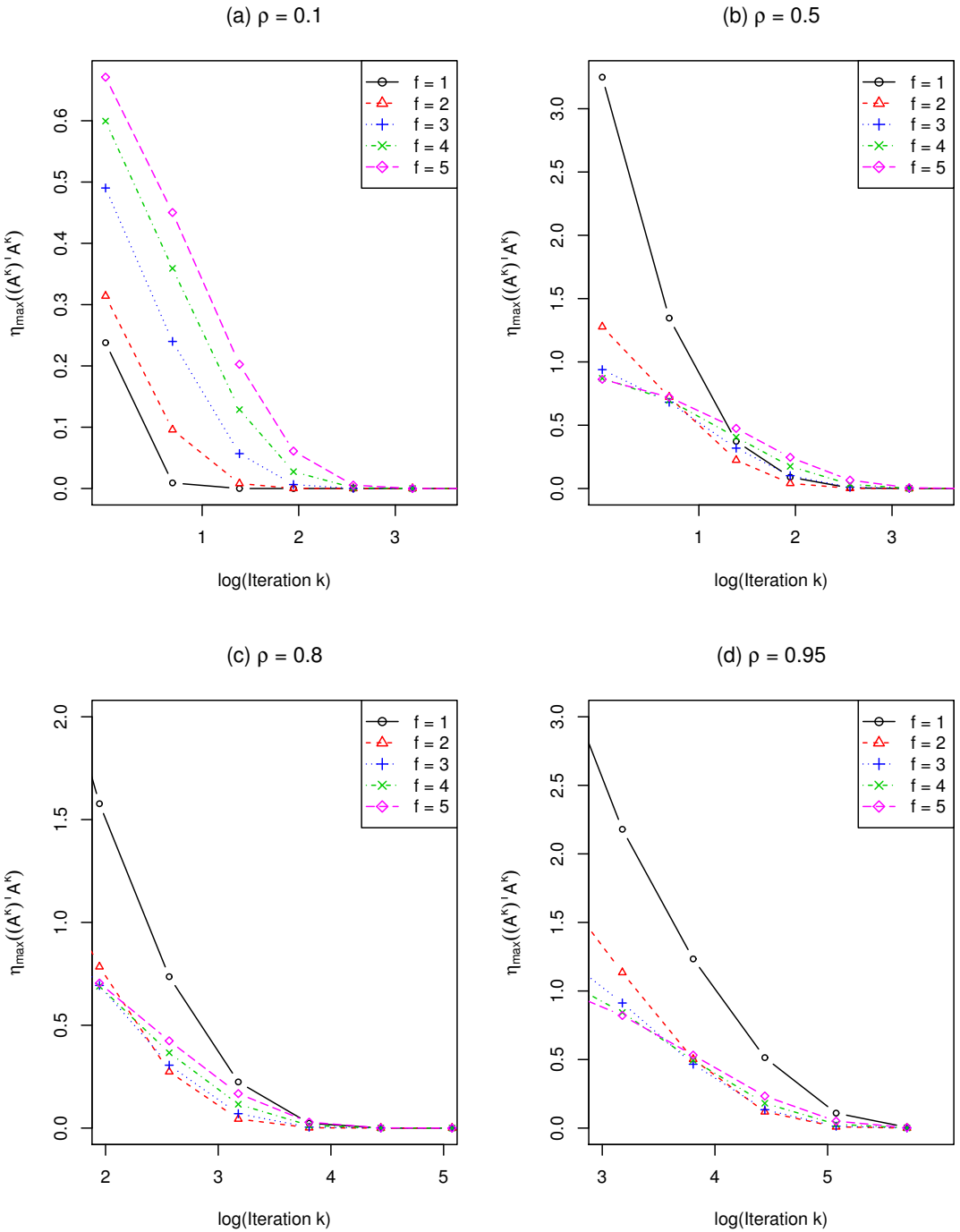
$$\mathbf{W}_j = \mathbf{I}_{p\times p} + \begin{bmatrix} \mathbf{0}_{p\times(j-1)} & \mathbf{u}_j & \mathbf{0}_{p\times(N-j)} \end{bmatrix}$$

$$\mathbf{A} = \prod_{j=1}^{p} \mathbf{W}_j \qquad \mu = \sum_{s=1}^{p-1} \left(v_s \prod_{j=s+1}^{p} \mathbf{W}_j \right) + v_p$$

If apply the CD and CMD to the LS problem, after a complete cycle from $j = 1$ to $j = p$, we get

$$\tilde{\beta}^{(k)} = \tilde{\beta}^{(k-1)} \mathbf{A} + \mu$$

The convergence rate is basically the maximum eigenvalue of $\left(\mathbf{A}^k\right)^\top \mathbf{A}^k$, which is affected by both f and ρ.


(a) ρ = 0.1
(b) ρ = 0.5
(c) ρ = 0.8
(d) ρ = 0.95
f = 1
f = 2
f = 3
f = 4
f = 5
log(Iteration k)
$\eta_{max}((A^K)^t A^K)$

	Colon	Prostate	WBCD	Ionosphere	Sonar
N	62	102	569	351	208
p	2000	6033	495 (30)	560 (32)	1890 (60)
α_{CV}	0.6	0.5	0.6	0.4	0.4
Test Error	8.3%	5%	1.77%	2.86%	24.39%
glmnet	0.1166	0.3283	9.4039	0.5158	2.0828
glmnet2	0.0910	0.2938	4.9593	0.3667	1.0945
Improv. %	+28%	+11.7%	+89.6%	+40.6%	+90.3%

ADMM

- Douglas & Rachford (1956); Lions & Mercier (1979); Eckstein & Bertsekas (1992)
- Goldstein & Osher (2009); Yin, Osher, Goldfard, and Darbon (2008); Goldfarb & Ma (2012)
- many applications in signal processing, statistics, machine learning

Improving MPT

Xue, Ma and Zou (2012)

An investor has p assets. Asset j makes up ω_j proportion of the investor's portfolio.

$$\omega_j \geq 0 \quad \sum_{j=1}^{p} \omega_j = 1.$$

Asset j delivers return R_j which has mean μ_i and variance σ_j^2. The mean of the return of the entire portfolio is $\sum_{j=1}^{p} \omega_j \mu_j$ and the variance of the portfolio's return is

$$\sum_{i}^{p} \sum_{j}^{p} \omega_i \omega_j \sigma_i \sigma_i \rho_{ij}$$

where ρ_{ij} is the correlation between R_i and R_j.

$\boldsymbol{w} = (\omega_1, \ldots, \omega_p)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$,
$\boldsymbol{\Sigma}$ is the covariance matrix of return vector $(R_1, \ldots, R_p)^{\mathrm{T}}$.

MPT

$$\underset{\boldsymbol{w}}{\operatorname{argmin}}\, \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{w}$$

$$\textit{s.t.}$$

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\mu} = \mu_P,\ \boldsymbol{w}^{\mathrm{T}}\vec{1} = 1\ ,$$

$$\omega_j \geq 0, j = 1, \ldots, p.$$

MPT (1952, J. of Finance) won 1990 Nobel Prize in Economics.

The usual implementation of MPT

$\hat{\boldsymbol{\mu}} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ is the sample mean vector

$\hat{\boldsymbol{\Sigma}}_n$ is the sample covariance matrix

Empirical MPT

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \boldsymbol{w}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{w}$$

s.t.

$$\boldsymbol{w}^{\mathrm{T}} \hat{\boldsymbol{\mu}} = \mu_P, \ \boldsymbol{w}^{\mathrm{T}} \vec{1} = 1 \ ,$$

$$\omega_j \geq 0, j = 1, \ldots, p.$$

When p is relatively large

- The sample cov. matrix performs poorly (Johnstone, 2001). It leads to bias and undesirable risk issues in the empirical MPT (El Karoui, 2010,Brodie et al., 2009; DeMiguel et al., 2009; Fan et al., 2012).
- Under some suitable "sparsity" assumption on $\mathbf{\Sigma}$, an optimal estimator can be obtained by Thresholding (Bickel and Levina 2008a; El Karoui, 2008,Cai and Zhou, 2011)
- Let $\hat{\sigma}_{ij}$ be the ij entry of the sample covariance matrix.

$$\hat{\mathbf{\Sigma}}_{thresholding} = \{s_\lambda(\hat{\sigma}_{ij})\}_{1\leq i,j\leq p}$$

The difficulty is how to preserve both P.D. and Sparsity simultaneously.

Notation: $|\mathbf{\Sigma}|_1 = \sum_{i \neq j} |\sigma_{ij}|$, $\|\mathbf{\Sigma}\|_F^2 = \sum_{i,j} \sigma_{ij}^2$.

The soft-thresholding estimator is the global solution of

$$\underset{\mathbf{\Sigma}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Sigma}|_1.$$

PSD sparse covariance estimator

$$\hat{\mathbf{\Sigma}}^+ = \underset{\mathbf{\Sigma} \succeq \epsilon \mathbf{I}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Sigma}|_1.$$

$\epsilon = 10^{-6}$. ϵ can be other positive constant depending on the application.

Algorithm

The augmented Lagrangian function for some given parameter μ ,

$$L(\mathbf{\Theta}, \mathbf{\Sigma}; \mathbf{\Lambda}) = \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Sigma}|_1 - \langle \mathbf{\Lambda}, \mathbf{\Theta} - \mathbf{\Sigma}\rangle + \frac{1}{2\mu}\|\mathbf{\Theta} - \mathbf{\Sigma}\|_F^2,$$

where $\mathbf{\Lambda}$ is the Lagrange multiplier.

For $i = 0, 1, 2, \ldots,$

$$\begin{aligned}
\mathbf{\Theta} \text{ step}: \quad \mathbf{\Theta}^{i+1} &= \arg\min_{\mathbf{\Theta} \succeq \epsilon \mathbf{I}} L(\mathbf{\Theta}, \mathbf{\Sigma}^i; \mathbf{\Lambda}^i) \\
\mathbf{\Sigma} \text{ step}: \quad \mathbf{\Sigma}^{i+1} &= \arg\min_{\mathbf{\Sigma}} L(\mathbf{\Theta}^{i+1}, \mathbf{\Sigma}; \mathbf{\Lambda}^i) \\
\mathbf{\Lambda} \text{ step}: \quad \mathbf{\Lambda}^{i+1} &= \mathbf{\Lambda}^i - \frac{1}{\mu}(\mathbf{\Theta}^{i+1} - \mathbf{\Sigma}^{i+1}).
\end{aligned}$$

Θ step

$$L(\mathbf{\Theta}, \mathbf{\Sigma}; \mathbf{\Lambda}) = \frac{1}{2}\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}_n\|_F^2 + \lambda|\mathbf{\Sigma}|_1 - \langle \mathbf{\Lambda}, \mathbf{\Theta} - \mathbf{\Sigma}\rangle + \frac{1}{2\mu}\|\mathbf{\Theta} - \mathbf{\Sigma}\|_F^2$$

$$\begin{aligned}
\mathbf{\Theta}^{i+1} &= \arg\min_{\mathbf{\Theta} \succeq \epsilon \mathbf{I}} L(\mathbf{\Theta}, \mathbf{\Sigma}^i; \mathbf{\Lambda}^i) \\
&= \arg\min_{\mathbf{\Theta} \succeq \epsilon \mathbf{I}} -\langle \mathbf{\Lambda}^i, \mathbf{\Theta}\rangle + \frac{1}{2\mu}\|\mathbf{\Theta} - \mathbf{\Sigma}^i\|_F^2 \\
&= \arg\min_{\mathbf{\Theta} \succeq \epsilon \mathbf{I}} \|\mathbf{\Theta} - (\mathbf{\Sigma}^i + \mu\mathbf{\Lambda}^i)\|_F^2 \\
&= (\mathbf{\Sigma}^i + \mu\mathbf{\Lambda}^i)_+.
\end{aligned}$$

Let $\mathbf{Z}$'s eigen-decomposition be $\sum_{j=1}^p \lambda_j v_j^{\mathrm{T}} v_j$, then define $(\mathbf{Z})_+ = \sum_{j=1}^p \max(\lambda_j, \epsilon) v_j^{\mathrm{T}} v_j$.

Σ step

$$L(\Theta, \Sigma; \Lambda) = \frac{1}{2}\|\Sigma - \hat{\Sigma}_n\|_F^2 + \lambda|\Sigma|_1 - \langle \Lambda, \Theta - \Sigma \rangle + \frac{1}{2\mu}\|\Theta - \Sigma\|_F^2$$

$$\begin{aligned}
\Sigma^{i+1} &= \arg\min_{\Sigma} L(\Theta^{i+1}, \Sigma; \Lambda^i) \\
&= \arg\min_{\Sigma} \frac{1}{2}\|\Sigma - \hat{\Sigma}_n\|_F^2 + \lambda|\Sigma|_1 + \langle \Lambda^i, \Sigma \rangle + \frac{1}{2\mu}\|\Sigma - \Theta^{i+1}\|_F^2 \\
&= \arg\min_{\Sigma} \frac{1}{2}\|\Sigma - \frac{\mu(\hat{\Sigma}_n - \Lambda^i) + \Theta^{i+1}}{1+\mu}\|_F^2 + \frac{\lambda\mu}{1+\mu}|\Sigma|_1 \\
&= \frac{1}{1+\mu}\mathbf{S}(\mu(\hat{\Sigma}_n - \Lambda^i) + \Theta^{i+1}, \lambda\mu).
\end{aligned}$$

Define $\mathbf{S}(\mathbf{Z}, \tau) = \{s(z_{j\ell}, \tau)\}_{1 \le j,\ell \le p}$ with
$s(z_{j\ell}, \tau) = \text{sign}(z_{j\ell}) \max(|z_{j\ell}| - \tau, 0) I_{\{j \neq \ell\}} + z_{j\ell} I_{\{j=\ell\}}$.

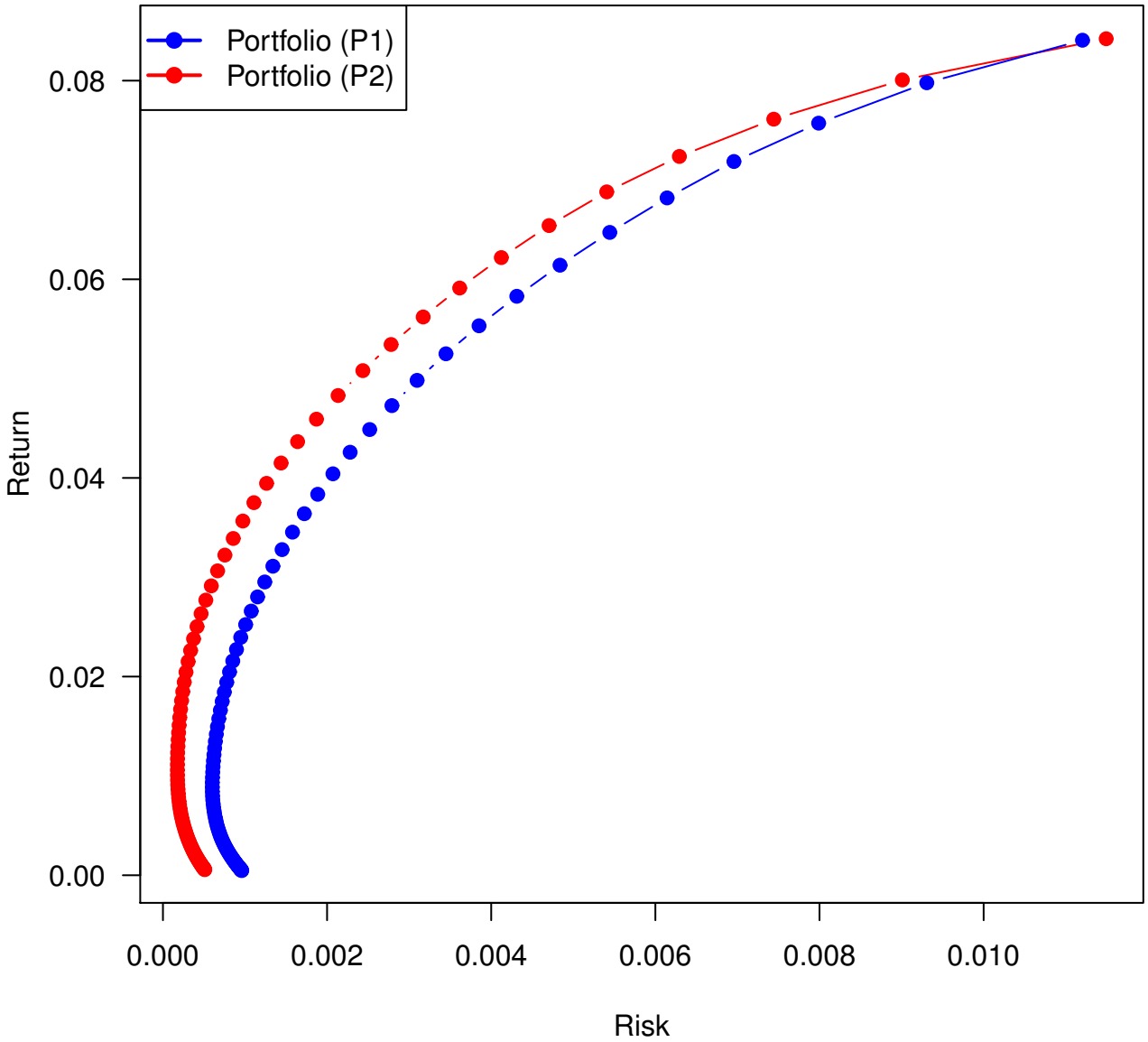
Improved Empirical MPT

$$\hat{w} = \arg\min_{w} w^{\mathrm{T}} \hat{\Sigma}^{+} w$$

$$s.t.$$

$$w^{\mathrm{T}} \hat{\mu} = \mu_P,\ w^{\mathrm{T}} \vec{1} = 1\ ,$$

$$\omega_j \geq 0, j = 1, \ldots, p.$$



Two MPT frontiers based on S&P 100 from Jan. 1990—Jan. 1993. Red: new;blue: traditional.

Latent Variable glasso

Ma, Xue and Zou (2013)

Latent variable Gaussian graphical model

- observed $\mathbf{X}$ (p-dim.) and unobserved $\mathbf{Y}$ (q-dim.) are jointly Gaussian

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N_{p+q}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_Y \end{bmatrix} \right)$$

- sparsity: $\mathbf{X}|\mathbf{Y}$ has a sparse Gaussian graphical model representation.
- $$(\boldsymbol{\Sigma})^{-1} = \boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_X & \boldsymbol{\Theta}_{XY} \\ \boldsymbol{\Theta}_{YX} & \boldsymbol{\Theta}_Y \end{bmatrix}$$

  $\mathbf{X}|\mathbf{Y}$ is normal with precision matrix $\boldsymbol{\Theta}_X$.
- How to estimate $\boldsymbol{\Theta}_X$ just based on $\mathbf{X}$?

A convex formulation

Chandrasekaran, Parrilo & Willsky (2012)

A key observation: $\mathbf{\Sigma}_X^{-1} = \mathbf{\Theta}_X - \mathbf{\Theta}_{XY}\mathbf{\Theta}_Y^{-1}\mathbf{\Theta}_{YX}$

- $\mathbf{\Theta}_X$ is sparse (assumption),
- $\mathbf{\Theta}_{XY}$ has rank at most q, $\mathbf{\Theta}_{XY}\mathbf{\Theta}_Y^{-1}\mathbf{\Theta}_{YX}$' rank is at most q.

If assume q is small (very reasonable in applications), we have a "sparse" -"low-rank" decomposition of $\mathbf{\Sigma}_X^{-1}$–the marginal precision matrix of $\mathbf{X}$.

LVGM estimator

Write

$$\mathbf{\Sigma}_X^{-1} = \mathbf{S} - \mathbf{L},$$

$\mathbf{S}$ is a sparse PD matrix and $\mathbf{L}$ is a low rank SPD matrix.

$$\begin{aligned} \min_{(\mathbf{S},\mathbf{L})} \quad & \langle \hat{\mathbf{\Sigma}}_X, \mathbf{S} - \mathbf{L} \rangle - \log\det(\mathbf{S} - \mathbf{L}) + \alpha \|\mathbf{S}\|_1 + \beta \mathrm{Tr}(\mathbf{L}) \\ \text{subject to} \quad & \mathbf{S} - \mathbf{L} \succ 0, \mathbf{L} \succeq 0 \end{aligned}$$

$\|\mathbf{S}\|_1$ is a convex relaxation of the sparsity of $\mathbf{S}$. $\mathrm{Tr}(\mathbf{L})$ is a convex relaxation of the rank of $\mathbf{L}$.

Chandrasekaran, Parrilo & Willsky (2012) viewed the above as a log-determinant semidefinite programming problem.

Algorithm

- $\mathbf{R} = \mathbf{S} - \mathbf{L}$

$$\begin{array}{ll} \min_{(\mathbf{R},\mathbf{S},\mathbf{L})} & \langle \hat{\mathbf{\Sigma}}_X, \mathbf{R} \rangle - \log\det(\mathbf{R}) + \alpha \|\mathbf{S}\|_1 + \beta \mathsf{Tr}(\mathbf{L}) \\ \text{subject to} & \mathbf{R} \succ 0,\ \mathbf{L} \succeq 0 \end{array}$$

- augmented Lagrangian

$$\begin{aligned} \mathcal{L}(\mathbf{R}, \mathbf{S}, \mathbf{L}; \mathbf{\Lambda}) \quad = \quad & \langle \hat{\mathbf{\Sigma}}_X, \mathbf{R} \rangle - \log\det(\mathbf{R}) + \alpha \|\mathbf{S}\|_1 + \beta \mathsf{Tr}(\mathbf{L}) \\ & - \langle \mathbf{\Lambda}, \mathbf{R} - \mathbf{S} + \mathbf{L} \rangle + \frac{1}{2\mu} \|\mathbf{R} - \mathbf{S} + \mathbf{L}\|_F^2. \end{aligned}$$

- alternating minimization

$$\begin{cases} \mathbf{R}^{k+1} & = & \arg\min_{\mathbf{R}} \ \mathcal{L}(\mathbf{R}, \mathbf{S}^k, \mathbf{L}^k; \mathbf{\Lambda}^k) \\ \mathbf{S}^{k+1} & = & \arg\min_{\mathbf{S}} \ \mathcal{L}(\mathbf{R}^{k+1}, \mathbf{S}, \mathbf{L}^k; \mathbf{\Lambda}^k) \\ \mathbf{L}^{k+1} & = & \arg\min_{\mathbf{L} \succeq 0} \ \mathcal{L}(\mathbf{R}^{k+1}, \mathbf{S}^{k+1}, L; \mathbf{\Lambda}^k) \\ \mathbf{\Lambda}^{k+1} & = & \mathbf{\Lambda}^k - \frac{1}{\mu}(\mathbf{R}^{k+1} - \mathbf{S}^{k+1} + \mathbf{L}^{k+1}) \end{cases}$$

R step

$$\arg\min_{\mathbf{R}} \langle \hat{\mathbf{\Sigma}}_X, \mathbf{R} \rangle - \log\det(\mathbf{R}) - \langle \mathbf{\Lambda}^k, \mathbf{R} - \mathbf{S}^k + \mathbf{L}^k \rangle + \frac{1}{2\mu} \|\mathbf{R} - \mathbf{S}^k + \mathbf{L}^k\|_F^2$$

$$\arg\min_{\mathbf{R}} - \log\det(\mathbf{R}) + \frac{1}{2\mu} \|\mathbf{R} - \mathbf{G}\|_F^2$$

$$\mathbf{G} = \mathbf{S}^k - \mathbf{L}^k - \mu(\hat{\mathbf{\Sigma}}_X - \mathbf{\Lambda}^k)$$

$$\mathbf{R} - \mathbf{G} - \mu\mathbf{R}^{-1} = 0$$

Let $\mathbf{G} = U^T \sigma U$ (eigen-decomposition of $\mathbf{G}$)

$$\mathbf{R} = U^T \gamma U$$

with

$$\gamma_i = \frac{\sigma_i + \sqrt{\sigma_i^2 + 4\mu}}{2}$$

S step

$$\arg\min_{\mathbf{S}} \alpha\|\mathbf{S}\|_1 - \langle \mathbf{\Lambda}^k, \mathbf{R}^{k+1} - \mathbf{S} + \mathbf{L}^k \rangle + \frac{1}{2\mu}\|\mathbf{R}^{k+1} - \mathbf{S} + \mathbf{L}^k\|_F^2$$

$$\mathbf{S}^{k+1} = \arg\min_{\mathbf{S}} \mu\alpha\|\mathbf{S}\|_1 + \frac{1}{2}\|\mathbf{Z} - \mathbf{S}\|_F^2$$

$$\mathbf{Z} = (\mathbf{R}^{k+1} + \mathbf{L}^k - \mu\mathbf{\Lambda}^k)$$

$\tau = \mu\alpha$

$$\mathbf{S}_{ij}^{k+1} = [Shrink(\mathbf{Z}, \tau)]_{ij} := \begin{cases} Z_{ii}, & \text{if } i = j \\ Z_{ij} - \tau, & \text{if } i \neq j \text{ and } Z_{ij} > \tau \\ Z_{ij} + \tau, & \text{if } i \neq j \text{ and } Z_{ij} < -\tau \\ 0, & \text{if } i \neq j \text{ and } -\tau \leq Z_{ij} \leq \tau. \end{cases}$$

L step

The above is equivalent to

$$\arg\min_{\mathbf{L}\succeq 0} \beta \mathrm{Tr}(\mathbf{L}) - \langle \mathbf{\Lambda}^k, \mathbf{R}^{k+1} - \mathbf{S}^{k+1} + \mathbf{L} \rangle + \frac{1}{2\mu} \|\mathbf{R}^{k+1} - \mathbf{S}^{k+1} + \mathbf{L}\|_F^2$$

$$\mathbf{L}^{k+1} = \arg\min_{\mathbf{L}\succeq 0} (\mu\beta)\mathrm{Tr}(\mathbf{L}) + \frac{1}{2}\|\mathbf{M} - \mathbf{L}\|_F^2$$

where

$$\mathbf{M} = (\mathbf{S}^{k+1} - \mathbf{R}^{k+1} + \mu\mathbf{\Lambda}^k)$$

$\mathbf{M} = U^T \sigma U$ (eigen-decomposition of $\mathbf{M}$) then

$$\mathbf{L}^{k+1} = SVT(\mathbf{M}, \mu\beta) = U^T \gamma U$$

with

$$\gamma_i = \max(\sigma_i - \mu\beta, 0)$$

Concluding remark

- Tailoring optimization algorithms to the specific statistics problem
- Efforts to polish the solver

Thank You