Topics in Variable Selection II

Penalized Likelihood and Sure Independence Screening

Elvis Cui

PhD in Biostatistics, UCLA

November 28, 2019

Overview

1. Intro to high dimensional problems
2. Classical model selection
   - L_0 penalty
   - AIC as L_0 penalty
   - BIC as L_0 penalty
   - Adjusted R^2 as L_0 penalty
3. Penalized likelihood and regularizations
   - Three criterions for a good estimator
   - L_q reg. and Bayes prior
   - Extensions of LASSO
   - Graphical models: Glasso and CLIME
   - SCAD-type penalties
   - Embarrassingly simple approach to zero-shot learning (ESAZSL)
   - Matrix regularizations: Spectrum, Rank-reduction, Operator norm

Intro to high dimensional problems

- Difficulties when $d \gg n$
  1. Collinearity between covariates
  2. Noise accumulation
- 2 types of statistical endeavors
  1. Accuracy of estimated parameters
  2. Accuracy of $\mathbb{E}L(\mu, \hat{\mu})$
- Solution
  - Dimension reduction
  - Regularization
- Ultra-high dimension problems
  1. $\log p = O(n^\alpha)$
  2. For instance, thousands of samples with millions of genes

Classical model selection: L_0 penalty

A unified approach to model selection:

$$L_n(\theta) - \lambda \|\theta\|_0$$

Where $L_n(\theta) = \log \mathbb{P}(X_1, \cdots, X_n|\theta)$ and $\|\theta\|_0 = \sum_{i=1}^{d} I\{\theta_i \neq 0\}$.

- This is a problem with NP-complexity.
- Many classical model selection rules can be viewed as a special case of this L_0 penalty.

L_0 penalty: examples

Example 1 (AIC)

Akaike proposed to minimize KL divergence between the true model and the fitted model:

$$\hat{\theta} = \arg\min_{\theta} KL(true \| fit)$$

The KL divergence can be approximated by (up to a constant)

$$-L_n(\hat{\theta}_{MLE}) + \lambda dim(\hat{\theta}_{MLE}) = -L_n(\hat{\theta}_{MLE}) + \lambda \sum_{i=1}^{d} I(-L_n(\hat{\theta}_j \neq 0))$$

This is the L_0 penalty with $\lambda = 1$.

L_0 penalty: examples

Example 2 (BIC)

In 1978, Schwarz proposed BIC:

$$-L_n(\theta) - \frac{\log n}{2} \sum_{j=1}^{d} I(\theta_j \neq 0)$$

This is the L_0 penalty with $\lambda = \frac{\log n}{2}$.

Example 3 (Mallows' C_p)

The Mallows' C_p estimate is given by

$$\frac{SSE_d}{n} + \frac{2\hat{\sigma}^2}{n} p$$

If we are willing to add normal assumption, this is equivalent to the L_0 penalty with $\lambda = 1$.

L_0 penalty: examples

Example 4 (Adjusted R^2)

The adjusted R^2 is defined to be

$$R^2_{adj} = 1 - \frac{n-1}{n-d}\frac{SSE_d}{SST}$$

This $\max R^2_{adj}$ is equivalent to $\min n \log \frac{SSE_d}{n-d}$. But we know that

$$\frac{SSE_d}{n-d} \approx \sigma^2$$

Thus, we have

$$n \log \frac{SSE_d}{n-d} \approx \frac{SSE_d}{\sigma^2} + d + n(\log \sigma^2 - 1)$$

Under normal assumption, this is the L_0 penalty with $\lambda = \frac{1}{2}$.

Penalized likelihood

- A general framework: the loss function is defined to be

$$Loss(\theta) = \frac{L_n(\theta)}{n} - \sum_{j=1}^{d} p_\lambda(|\theta_j|)$$

  where the first term is log-likelihood and the second term $p_\lambda(|\theta_j|)$ is the penalty term for each component θ_j depending on λ.
- This was proposed by J. Fan and R. Li in 2001.
- Sometimes, we write the penalty term in a more compact way: $p_\lambda(\theta)$
- Lots of examples will be given in the following.

Criterions for a good estimator $\hat{\theta}$

- **Sparsity** (Fan and Li, 2001): $\hat{\theta}$ should be sparse so that it does feature selection automatically.
- **Unbiasedness** (Fan and Li, 2001): $\hat{\theta}$ should be approximated unbiased.
- **Continuity** (Breiman, 1996): $\hat{\theta}$ should be continuous w.r.t. λ, X and y so that the interpretability is strong.
- Note that subset selection algorithms are discrete since each variable is either selected or discarded.

Examples of penalty functions

Example 5 (Ride regression)

[Hoerl and Kennard, 1970]

$$p_\lambda(|\theta_j|) = \lambda\theta_j^2$$

or equivalently

$$\hat{\theta} = \arg\min -\frac{L_n(\theta)}{n} \text{ s.t. } \|\theta\|_2^2 \leq t$$

Note that for least square (or normal assumption), the degrees of freedom of the model is

$$df(\lambda) = Tr(X(X^TX + \lambda I)^{-1}X^T) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Where d_j is the singular value of X.

Examples of penalty functions

Example 6 (LASSO)

[Tibshirani, 1996] This is $p_\lambda(|\theta_j|) = \lambda|\theta_j|$ so that

$$p_\lambda(\theta) = \lambda\|\theta\|_1$$

Example 7 (L_q penalty as Bayes' estimator)

If we take $p_\lambda(|\theta_j|) = \lambda|\theta|^q$, then for each q, this corresponds to a Bayes' prior on θ:

$$\log \mathbb{P}(\theta) = \lambda \sum_{j=1}^{p} |\theta_j|^q + C_0$$

- q=1, this is Laplace prior (**Biased estimator**).
- q=2, Gaussian prior (**not sparse**).
- q¡1, concave penalty (**not continuous**).

Examples of penalty functions

Example 8 (Elastic net)

[Zou and Hastie, 2005] If we want both sparsity and shrinkage estimation, then we could take

$$p_\lambda(\theta) = \lambda \sum_{j=1}^{d} (\alpha\theta_j^2 + (1-\alpha)|\theta_j|)$$

Such penalty has **computational advantages** over L_q penalties.

Examples of penalty functions

Example 9 (Group LASSO)

[Bakin 1999; Lin and Zhang, 2006] If we want a group of parameters to vary simultaneously (which is quite often in biology), we could take

$$p_\lambda(\theta) = \lambda \sum_{l=1}^{L} \sqrt{p_l} \|\theta_l\|_2$$

where $\theta_l = (\theta_{l1}, \cdots, \theta_{lt})^T$ is the l^{th} group and $\theta = (\theta_1, \cdots, \theta_L)^T$.

Example 10 (General Group LASSO)

[Zhao et al, 2008] We take another norm:

$$\|\theta_l\|_K = (\theta_l^T K \theta_l)^{\frac{1}{2}}$$

and allow overlapping between groups.

Examples of penalty functions

Example 11 (Graphical LASSO)

In a undirected Gaussian graphical model, $\Theta \in \mathbb{R}^{d \times d}$ is the precision matrix, and we put penalty

$$p_\lambda(\Theta) = \|\Theta\|_{L_1,\mathrm{off}}$$

to encourage sparsity in Θ, which is conditional independence ($\|\cdot\|_{L_1,\mathrm{off}}$ denotes the elementwise L_1 norm except for diagonals).

A digression on CLIME

Example 12 (CLIME)

[Cai et al, 2012] (A constrained l_1 minimization approach to sparse precision matrix estimation) Let $\{\widehat{\Theta}_1\}$ be the solution set of the following optimization problem:

$$\min \|\Theta\|_1 \text{ subject to:} \tag{1}$$

$$\|\Sigma_n \Theta - I\|_{\max} \leq \lambda_n, \Theta \in \mathcal{R}^{d\times d} \tag{2}$$

Note that the solution $\widehat{\Theta}_1$ may not be symmetric. Thus if we write $\widehat{\Theta}_1 = (\hat{w}_{ij}^1)$, then the CLIME estimator $\widehat{\Theta}_{\text{CLIME}}$ of Θ^* is defined by one-step symmetrization:

$$\widehat{\Theta}_{\text{CLIME}} = (\hat{w}_{ij}), \tag{3}$$

$$\text{where } \hat{w}_{ij} = \hat{w}_{ji} = \hat{w}_{ij}^1 I\{|\hat{w}_{ij}^1 \leq \hat{w}_{ji}^1|\} + \hat{w}_{ji}^1 I\{\hat{w}_{ij}^1 > \hat{w}_{ji}^1\} \tag{4}$$

Examples of penalty functions

Example 13 (SCAD)

[Fan and Li, 2001] (Smoothly Clipped Absolute Deviation) If we want **sparsity, unbiasedness and continuity**, then we could take the derivative of $p_\lambda(t)$ to be $(t = |\theta_j|)$

$$p'_\lambda(t) = \lambda\{I(t \leq \lambda) + \frac{(\alpha\lambda - t)_+}{(\alpha - 1)\lambda} I(t > \lambda)\}, \ \alpha > 2$$

Integrating gives

$$p_\lambda(t) = \lambda t I(t \leq \lambda) + \frac{(a\lambda t - \frac{t^2+\lambda^2}{2})}{(a-1)} I(\lambda < t \leq \alpha\lambda) + \frac{\lambda^2\alpha^2}{2(\alpha - 1)} I(t \geq \alpha\lambda)$$

Note that this is a non-convex/non-concave penalty and SCAD coincides with LASSO if $|\theta_j| \leq \lambda$.

Examples of penalty functions

Example 14 (MCP)

[Minimax Concave Penalty] Similar to SCAD,

$$p_\lambda(|\theta_j|) = (\lambda|\theta_j| - \frac{\theta_j^2}{2\alpha})I(|\theta_j| \leq \alpha\lambda) + \frac{\alpha\lambda^2}{2}I(|\theta_j| > \alpha\lambda),\ \lambda > 1$$

Its derivative is

$$p'_\lambda(t) = \frac{(\alpha\lambda - t)_+}{\alpha}$$

Example 15 (Hard-threshold)

The penalty is given by $p_\lambda(|\theta_j|) = \lambda^2 - (\lambda - |\theta_j|)_+^2$. If we assume X has orthogonal columns, then there is an analytical solution of hard-threshold estimator.

A digression on computer vision

Example 16 (Embarrassingly simple approach to zero-shot learning)

[Romera-Paredes and Torr, 2015] Let $X \in \mathbb{R}^{d \times m}$, $S \in [0,1]^{a \times z}$ and $Y \in \{0,1\}^{m \times z}$ (ground truth label).

- d: dimension of each image
- m: number of instances
- z: number of classes
- a: number of attributes

The ESAZL estimator is defined to be

$$\widehat{V} = \arg \min_{V \in \mathbb{R}^{d \times a}} Loss(X^T VS, Y) + p_\lambda(V)$$

Where $Loss(\cdot, \cdot)$ can be taken as hinge loss, Frobenious loss, etc. And $p_\lambda(V)$ is the regularization term.

ESAZSL cont'd

Example 17 (Closed form solution)

Now if we take

$$Loss(A, B) = \|A - B\|_{Fro}^2 \tag{5}$$

$$p_\lambda(V) = \lambda_1 \|VS\|_{Fro}^2 + \lambda_2 \|X^T V\|_{Fro}^2 + \lambda_3 \|V\|_{Fro}^2 \tag{6}$$

$$\lambda_3 = \lambda_1 \lambda_2 \tag{7}$$

Then we have a closed form solution of ESAZSL:

$$\widehat{V} = (XX^T + \lambda_1 I)^{-1} XYS^T (SS^T + \lambda_2 I)^{-1}$$

Matrix regularizations

Example 18 (Matrix regularizations)

We have seen CLIME and Glasso in the previous slides. However, to encourage different types of sparsity:

- Low-rank
- Low radius of spectrum
- Sparse principal components
- Low sum of singular values

We need different types of matrix regularizations, see **Hua Zhou**'s 2014 paper on matrix regularizations.

END !

Next time: MCMC, Tree-based algorithms and Boosting