p-values in high-dimensional regression

Kaspar Rufibach

Biostatistics Unit
Institute of Social and Preventive Medicine
University of Zurich

Zurich, December 3rd 2009

Content

Setup

High-dimensional linear regression, $p \gg n$.

Questions:

- Variable selection.
- Assess uncertainty.
- Assign significance.

Lasso

Variable selection method, works for $p \gg n$:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p} \left(n^{-1} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right)$$

$\lambda > 0$: penalty parameter, typically chosen via CV.

Properties of Lasso:

- $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$: shrunken LS-estimate.
- Variable selection: some of the $\widehat{\beta}_j(\lambda)$ are exactly 0. Reason: "ℓ_1 geometry" (dual formulation).
- No significance testing: only get estimate via convex optimization.

Problem: Lasso typically selects "too many variables".

Asymptotics for Lasso

Lasso is consistent: Define

$$\widehat{\mathcal{S}}(\lambda) \quad = \quad \{j \ : \ \widehat{\beta}_j(\lambda) \neq 0\}$$

and assume

- sufficient and necessary neighborhood stability condition on $\mathbf{X}$,
- $p_n = O(n^\alpha)$ for some $0 < \alpha < \infty$ (high-dimensionality),
- $\mathcal{S}_{\mathrm{true},n}(\lambda) \ = \ O(n^\kappa)$ for some $0 < \kappa < 1$ (sparsity),
- non-zero β_j are "larger" than $n^{-1/2}$,
- $\mathbf{Y}, \mathbf{X}^j$ normal (can be generalized),
- $\lambda = \lambda_n = O(n^{1/2+\delta/2})$ for $0 < \delta < 1/2, \delta = \delta(\alpha, \kappa)$.

Then:

$$\begin{aligned} P(\widehat{\mathcal{S}}(\lambda) = \mathcal{S}_{\mathrm{true}}) \quad &= \quad 1 - O(exp(-Cn^{1-\delta})) \\ &\approx \quad 1 \text{ even for relatively small } n. \end{aligned}$$

Asymptotics for Lasso

"Lasso selects too many variables" $\Leftrightarrow$ consistency ?!

Problem 1: Is theory misleading? No. But conditions for consistency are

- restrictive and
- can not be checked for real data.

$\Rightarrow$ Lasso often not consistent for selecting relevant variables.

Problem 2: choice of λ. Even for prediction oracle solution Lasso yields asymptotically too large models!

Problem 3: Small non-zero coefficients can not reliably be detected.

Lasso entirely useless?

Lasso useless? No.

- With high probability: $\mathcal{S}_{\text{true}} \subseteq \widehat{\mathcal{S}}$.
- Typically for prediction optimal λ_{opt}: $\mathcal{S}_{\text{true}} \subseteq \widehat{\mathcal{S}}$.
- $\#\widehat{\mathcal{S}} \; = \; O(n)$ if $p \gg n \Rightarrow$ huge dimensionality reduction.

$\Rightarrow$ use Lasso as screening procedure.

2-stage recipe (Medallion lecture PB, 2009):

- "Pre-screen" variables with Lasso,
- OLS on $\widehat{\mathcal{S}}$ with variable selection (AIC, BIC).

But:

- May be unstable,
- no measure of significance.

Recipe to control FWER

p-values for the null hypothesis: $H_0^j \ : \beta_j \ = \ 0$.
$\Rightarrow$ generate p-value distribution via multi-sample splitting.

FWER: family-wise error rate, probability of falsely rejecting ≥ 1 null hypothesis. Strong sense, i.e. independent of alternative configuration.

Multi-split method:

- Let $\tilde{\mathcal{S}}$ be screening procedure that estimates $\widehat{\mathcal{S}}$.
- Do the following $b = 1, \ldots, B$ times:
  - Randomly split data in D_{in}^b and D_{out}^b,
  - use $\tilde{\mathcal{S}}$ to estimate $\widehat{\mathcal{S}}^b$,
  - fit selected variables $\widehat{\mathcal{S}}^b$ to D_{out}^b (e.g. via OLS), compute corresponding valid p-values $\tilde{p}_j^b$,
  - for variables not in $\widehat{\mathcal{S}}^b$ set $\tilde{p}_j^b = 1$.
- (Bonferroni-)Adjust (not yet aggregated) p-values:

$$P_j^b \quad = \quad \min\Big(\tilde{p}_j^b \cdot \#\tilde{S}^b, 1\Big), j = 1, \ldots, p.$$

Aggregation of p-values

Be aware. p-values are dependent!

Suitable aggregation of B p-values via quantiles:

- Compute for $\gamma \in (0, 1)$ and q_γ emp. quantile function:

$$Q_j(\gamma) \quad = \quad \min\Big\{1, q_\gamma(\{P_j^b/\gamma, b = 1, \ldots, B\})\Big\}.$$

- For fixed γ: Q_j is multiplicity adjusted asymptotically correct p-value.
- Choice of γ: error control not guaranteed if we search over γ. Slightly modify Q_j, use $\gamma_{\min} = 0.05$.

Theorem: these p-values control FWER. Similar procedure for (less conservative) FDR.

Alternative to p-values: Stability selection (Meinshausen & Bühlmann, 2009). Combine sub-sampling with high-dimensional selection procedures. May be consistent even if selection procedure is not.

Assumptions on "pre-screening" method

- Gaussian linear model with fixed design.
- Screening property: $\lim_{n\to\infty} P(\mathcal{S} \subseteq \tilde{\mathcal{S}}) = 1$.
  Relevant variables are retained.
- Sparsity property: $\#\tilde{S} < n/2$. Not too many noise variables are allowed to be selected (ensures validity of tests).

Assumptions are satisfied for:

- Lasso,
- L_2 boosting,
- sure independence screening,
- PC-algorithm in graphical models,
- etc.

Example: model age with plasma markers

Goals:

- Find plasma markers that are associated with age,
- knock-out corresponding genes in mice and worms.

Data of 184 healthy probands, 90 markers + sex.

Lasso estimate: `R` package `penalized`. Very easy to use.

Lasso pre-screening, OLS with BIC model selection.

Results: see consulting report.