

Asymptotic properties for some Lasso-type estimators: Parametric case



Xing Wang 王星

2008.06

School of Statistics

Renmin University Of China

Joint work with Guilherme & Binyu

wangxingscy@gmail.com

**2008 Beijing International Conference On
Machine Learning And Data Mining**

Two development in statistical machine learning

- With machine learning algorithms to get better solution for statistical problems
 - Efficient algorithms design (Boosting-type)
 - Sparsity pursue (Lasso-type)
 - Predictive performance improvement (SVM-type)
- Giving statistical insights on learning algorithms based above----asymptotic is one of the interest

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- Consistency simulation and results
- Conclusions

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- Consistency simulation and results
- Conclusions

Statistical model selection

- The parameters of interest in statistical models often represent the attempt to make optimal choices in the presence of uncertainty.
- Most parameters are defined as the minimizer of a *risk function* defined as the expected value of a loss function:

$$E(L(z, b)) = \int L(z, b) dP(z)$$

$$\beta = \arg \min_b E(L(z, b))$$

Penalization Framework

- Given data $Z_i = (Y_i, X_i)$, $i = 1, \dots, n$:
 - X_i : a p -dimensional predictor
 - Y_i : response variable
- The parameters β are defined by the penalized problem:

$$\hat{\beta}(\lambda) = \arg \min \left\{ \sum_{i=1}^n L(Z_i; \beta) + \lambda T(\beta) \right\}.$$

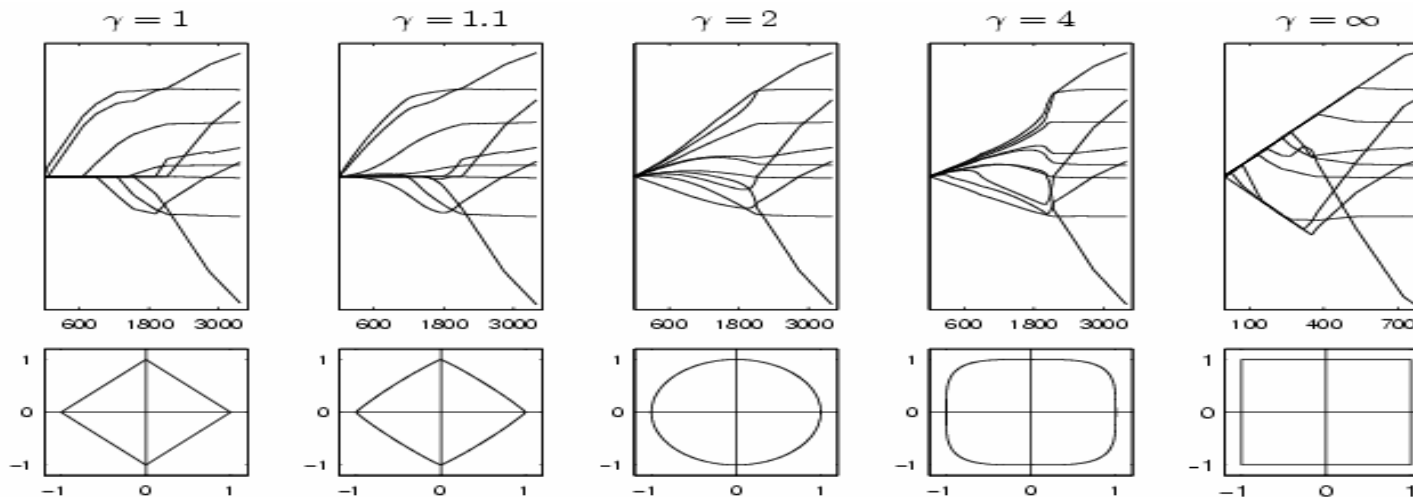
where

- $\sum_i L(Z_i; \beta)$ is the empirical loss function
- $T(\beta)$ is a penalty function
- λ is a tuning parameter

Types of penalties usually used

- l_γ -norm of the fitted coefficients defined as:

$$\|b\|_\gamma := \left(\sum_j |b_j|^\gamma \right)^{\frac{1}{\gamma}} \quad (\text{Frank Friedman, 1993})$$



Composite Absolute Penalties

- The CAP parameter estimate is given by:

$$\hat{\beta}(\lambda) = \arg \min_{\beta} L(Z; \beta) + \lambda \left\| \left[\|\beta_{G_1}\|_{\gamma_1}, \|\beta_{G_2}\|_{\gamma_2}, \dots, \|\beta_{G_k}\|_{\gamma_k} \right] \right\|_{\gamma_0}$$

- G_k 's, $k=1, \dots, K$ - indices of k -th pre-defined group
- β_{G_k} – corresponding vector of coefficients.
- $\|\cdot\|_{\gamma_k}$ – group L_{γ_k} norm: $N_k = \|\beta_{\gamma_k}\|_{\gamma_k}$;
- $\|\cdot\|_{\gamma_0}$ – overall norm: $T(\beta) = \|M\|_{\gamma_0}$
- groups may overlap (hierarchical selection)

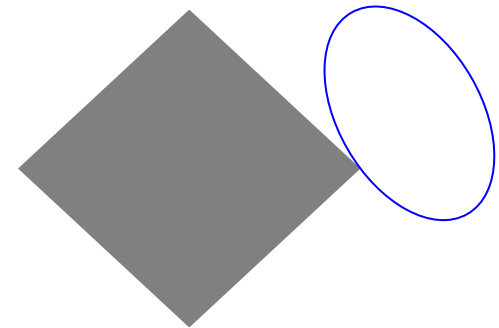
Lasso: L_1 -norm as a penalty

- The L_1 penalty is defined for coefficients β :

$$\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$$

- Used initially with L_2 loss:
 - Signal processing: Basis Pursuit (Chen & Donoho, 1994)
 - Statistics: LASSO (Tibshirani, 1996)

$$\hat{\beta} = \arg \min \left\{ \sum_{i=1}^n (Y_i - X_i \beta)^2 + \lambda \|\beta\|_1 \right\}.$$



- Properties:
 - Sparsity (variable selection)
 - Convexity (convex relaxation of L_0 -penalty)
 - Stability (non-negative garrote, Breiman, 1995)

L_1 penalty application and algorithms

- Cox regression
 - Tibshirani,1997
- Logistic regression,
 - Genkin,2007;
 - Zhu&Hastie,2004;
 - Park&Hastie,2006;
- Hinge loss
 - Zhu&Hastie,2004;
- Quantile Regression
 - Li&Zhu,2006.

New penalty functions

- Grouped LASSO
 - Yuan and Lin,2006
- Elastic Net
 - Zou and Hastie,2005
- Composite Absolute Penalties(CAP)
 - Zhao&YU2008;
- This conference.....

Consistency of Lasso

- Knight & Fu(2000) obtained results characterizing the behavior of the penalized least squares l_γ estimates in terms of a random function.
- Define:

$$Z_n(\varphi) = \frac{1}{n} \sum_{i=1}^n (Y_i - x_i^T \varphi)^2 + \frac{\lambda_n}{n} \sum_{j=1}^p |\varphi_j|^\gamma$$

which is minimized at $\varphi = \hat{\beta}_n$

Key theory in Knight& Fu

- Assume:

$$C = \frac{1}{n} XX^T \text{ is nonsingular;}$$

$$\lambda_n / n \rightarrow \lambda_0 \geq 0$$

- Main result:

$$Z(\varphi) = (\varphi - \beta)^T C(\varphi - \beta) + \lambda_0 \sum_{j=1}^p |\varphi_j|^\gamma$$

$$\lambda_n = o(n), \arg \min Z = \beta$$

$$\hat{\beta}_n \text{ is consistent, } \hat{\beta}_n \xrightarrow{p} \beta$$

Key theory in Knight& Fu

- Assume: $\gamma \geq 1$

$$C = \frac{1}{n} XX^T \text{ is nonsingular;}$$

$$\lambda_n / \sqrt{n} \rightarrow \lambda_0 \geq 0$$

- Main result:

$$\hat{\beta}_n \text{ is consistent, } \sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \arg \min(V)$$

$$V(u) = -2u^T W + u^T C u + \lambda_0 \sum_{j=1}^p u_j \text{sign}(\beta_j) |\beta_j|^\gamma$$

$$\text{When } \gamma = 1, W \sim N(0, \sigma^2 C)$$

Two Questions

- What's going on in asymptotics for more general estimators with other loss functions?
- How about asymptotics for a broad class of penalty functions?

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- Consistency simulation and results
- Conclusions

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- Consistency simulation and results
- Conclusions

Some assumptions

- We first define “re-centered” and “re-scaled” penalized objective function V involving both the risk and the penalty:

$$V_{\beta}^{(n)}(Q, \lambda, \mu) = n \left[QL\left(Z, \beta + \frac{\mu}{q_n}\right) - QL(Z, \beta) \right] + \lambda \left[T\left(\beta + \frac{\mu}{q_n}\right) - T(\beta) \right]$$

Lemma 1. *The minimizer of the re – centered and re – scaled objective function $V_{\beta}^{(n)}$:*

$$q_n (\hat{\beta}(\lambda_n) - \beta) = \arg \min_b V_{\beta}^{(n)}(\hat{P}_n, \lambda_n, b)$$

The key theorem

Theorem 2. Suppose there exist functions C_β, D_β and a random variable \mathbf{W} such that, for all compact sets $K \subset \mathbb{R}^d$:

$$i) \sup_{u \in K} \left| n \left[\hat{\mathbb{P}}_n L \left(\mathbf{Z}, \beta + \frac{\mathbf{u}}{q_n} \right) - \mathbb{P} L (\mathbf{Z}, \beta) \right] - C_\beta(\mathbf{W}, u) \right| \xrightarrow{P} 0;$$

$$ii) \sup_{u \in K} \left| \lambda_n \left[T \left(\beta + \frac{\mathbf{u}}{q_n} \right) - T (\beta) \right] - D_\beta(u) \right| \xrightarrow{P} 0;$$

Then, letting $V_\beta(\mathbf{W}, u) = C_\beta(\mathbf{W}, u) + D_\beta(u)$:

$$a) \sup_{u \in K} \left| V_\beta^{(n)}(\hat{\mathbb{P}}_n, \lambda_n, u) - V_\beta(\mathbf{W}, u) \right| \xrightarrow{P} 0;$$

If furthermore, the (un-penalized) $\hat{\beta}_n(0)$ is $O_p(1)$, then:

$$b) q_n \left(\hat{\beta}(\lambda_n) - \beta \right) \xrightarrow{d} \arg \min_u V_\beta(\mathbf{W}, u).$$

Sufficiency conditions for Loss function consistency

Assumption Set 1. *(Standard loss function assumptions)*

AL.I $\mathbb{P} |L(Z, b)| < \infty$ for each b ;

AL.II The loss function $L(Z, b)$ is such that:

- a) $L(Z, b)$ is differentiable in b at β for \mathbb{P} -almost every Z with derivative $\nabla_b L(Z, \beta)$
and:

$$\mathbb{P} [\nabla_b L(Z, \beta) \nabla_b L(Z, \beta)'] < \infty;$$

- b) the risk function $R(\mathbb{P}, b)$ is twice differentiable with respect to b at $b = \beta$ with positive definite Hessian matrix:

$$[\mathbf{H}(b)]_{ij} := \frac{\partial^2 R(\mathbb{P}, b)}{\partial \beta_i \partial \beta_j}.$$

AL.III The loss function $L(Z, b)$ is convex in b for \mathbb{P} -almost every Z ;

Loss convergence

Theorem 3 Under assumption AL.I and AL.II:

a) There exists $W \sim N(0, \Sigma)$ with $\Sigma = P[\nabla_b L(Z, \beta) \nabla_b L(Z, \beta)']$ such that :

$$P_n L(Z, \beta + \frac{u}{\sqrt{n}}) - PL(Z, \beta) - [u' \cdot H(\beta) \cdot u + W \cdot u] \xrightarrow{p} 0, \text{ for each } u \in \mathbb{R}^p;$$

b) If assumption AL.III holds, then for every compact subset $K \subset \mathbb{R}^p$,

$$\sup_{u \in K} \|\hat{P}_n L(Z, \beta + \frac{u}{\sqrt{n}}) - PL(Z, \beta) - [u' \cdot H(\beta) \cdot u + W \cdot u]\| \xrightarrow{p} 0;$$

Penalty convergence (Norm)

Lemma 7. (*Convergence of norm penalties*)

Suppose $T(b) = \|b\|_\gamma$ and $q_n \rightarrow \infty$. If

$$\gamma \in [1, \infty) \text{ with } \frac{\lambda_n^P}{q_n} \rightarrow \lambda_0 \text{ or}$$

$$\gamma \in [0, 1) \text{ with } \frac{\lambda_n^P}{q_n} \rightarrow \lambda_0,$$

then for all compact sets K :

$$\sup_{u \in K} \left\| \lambda_n \left[T\left(\beta + \frac{\mathbf{u}}{q_n}\right) - T(\beta) \right] - \lambda_0 G(\beta, \mathbf{u}) \right\| \xrightarrow{P} 0;$$

with :

$$G(\beta, \mathbf{u}) = \begin{cases} \sum_{j=1}^p u_j \text{sign}(\beta_j) |\beta_j|^{\gamma-1}, & \text{if } \gamma \in [1, \infty) \\ \sum_{j=1}^p [u_j \text{sign}(\beta_j) + |u_j| \mathbf{I}(\beta_j = 0)], & \text{if } \gamma = 1 \\ \sum_{j=1}^p |u_j|^\gamma \mathbf{I}(\beta_j = 0), & \text{if } \gamma \in (0, 1) \end{cases}$$

Types of Loss function usually used

- In regression problems,

$$L(Y, Xb) = (Y - Xb)^2$$

- In quantile regression,

$$L(Y, Xb) = \alpha |Y - Xb|_+ + (1 - \alpha) |Y - Xb|_-$$

- In classification,

– Logistic regression $L(Y, Xb) = \log(1 + \exp(-Yb'X))$

– Hinge loss $L(Y, Xb) = (1 - Yb'X)_+$

– Hüber robust loss $L(Y, Xb) = \begin{cases} |Y - Xb| / 2 & |Y - Xb| \leq \delta \\ \delta(|Y - Xb| - \delta / 2) & |Y - Xb| > \delta \end{cases}$

Example 1: Twice differentiable neg-loglikelihood functions

Corollary 4. (Uniform convergence for smooth neg – loglikelihood loss functions)
 suppose that the loss functions L is convex and twice continuously differentiable for almost all Z at a neighborhood of β with derivative $\nabla_b L(Z, \beta)$ and Hessian matrix $\frac{\nabla_b L(Z, \beta)}{\partial b_j \partial b_k}$, satisfying :

$$i) P \left| \frac{\nabla_b L(Z, \beta)}{\partial b_j \partial b_k} \right| < \infty \text{ for } b \text{ in a neighborhood of } \beta;$$

$$ii) P[\nabla_b L(Z, \beta) \nabla_b L(Z, \beta)'] < \infty$$

Then for all compact sets $K \subset \Theta \subset R^p$:

$$\sup_{u \in K} \left\| \hat{P}_n L(Z, \beta + \frac{u}{\sqrt{n}}) - PL(Z, \beta) - [u' \cdot H(\beta) \cdot u + W \cdot u] \right\| \xrightarrow{P} 0;$$

with

$$W \sim N(0, P[S(Z, \beta)S(Z, \beta)']), \text{ with } S(Z, \beta) = \frac{\partial L(Z, \beta)}{\partial b} \Big|_{b=\beta}$$

$$[H(\beta)]_{jk} = P \left[\frac{\nabla_b L(Z, \beta)}{\partial b_j \partial b_k} \Big|_{b=\beta} \right], \text{ for } j, k = 1, \dots, p$$

Example 2: Quantile regression loss functions

Assumption Set 2. (Quantile regression error assumption) For the quantile regression, we assume that:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where ε has a continuous distribution (w.r.t. Lebesgue measure) and $\mathbb{P}|\varepsilon| < \infty$. We let q_α denote the α -percentile of ε , i.e. q_α satisfies $\mathbb{P}[\varepsilon \leq q_\alpha] = \alpha$

Corollary 5. (Uniform convergence for quantile loss functions)

Suppose the loss function L is given $\mathbb{P}(\mathbf{X}\mathbf{X}') < \infty$ and the quantile regression error assumption is satisfied. Then, for all compact sets $K \subset \Theta \subset \mathbb{R}^p$:

$$\sup_{\mathbf{u} \in K} \left\| \hat{\mathbb{P}}_n L \left(\mathbf{Z}, \beta + \frac{\mathbf{u}}{\sqrt{n}} \right) - \mathbb{P}L(\mathbf{Z}, \beta) - [\mathbf{u}' \cdot \mathbf{H}(\beta) \cdot \mathbf{u} + \mathbf{W} \cdot \mathbf{u}] \right\| \xrightarrow{P} 0;$$

$$\mathbf{W} \sim N \left(0, \alpha \cdot (1 - \alpha) \cdot \mathbb{P}(\mathbf{X}\mathbf{X}') \right)$$

$$\mathbf{H}(\beta) = f(q_\alpha)^2 \mathbb{P}(\mathbf{X}\mathbf{X}')$$

Penalty convergence (CAP)

Lemma 8. (convergence of CAP penalties)

Suppose T as defined before, with $\gamma_0 = 1$ and $\gamma_k \in (1, \infty)$ for $k = 1, \dots, K$.

assume further that $q_n \rightarrow \infty$, $\frac{\lambda_n^P}{q_n} \rightarrow \lambda_0$ and $\beta_{G_k} \neq 0$ for all $k = 1, \dots, K$.

Then, for $\gamma \in (0, \infty)$ and for any compact set $K \subset \mathbb{R}^P$:

$$\sup_{u \in K} \left\| \lambda_n \left[T\left(\beta + \frac{\mathbf{u}}{\sqrt{n}}\right) - T(\beta) \right] - \lambda_0 G(\beta \cdot \mathbf{u}) \right\|^P \rightarrow 0;$$

where:

$$G(\beta \cdot \mathbf{u}) = \sum_{j=1}^p \sum_{k; j \in G_k} \left(\frac{u_j \text{sign}(\beta_j)}{\|\beta_{G_k}\|_{\gamma_k}} \right)^{\gamma_k - 1}$$

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- Consistency simulation and results
- Conclusions

Outline

- Some comments on asymptotic and consistency results.
- Consistency results in LASSO-type estimators.
- **Consistency simulation and results**
- Conclusions

Logistic Regression

$$P[Y | \mathbf{X}] = \frac{\exp(\alpha + \mathbf{X}\beta)}{1 + \exp(\alpha + \mathbf{X}\beta)}$$

To do that, we fix $p = 2$ and set :

$$\alpha = 0$$

$$\beta = (1, 0);$$

$$P(\mathbf{X}\mathbf{X}') = \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

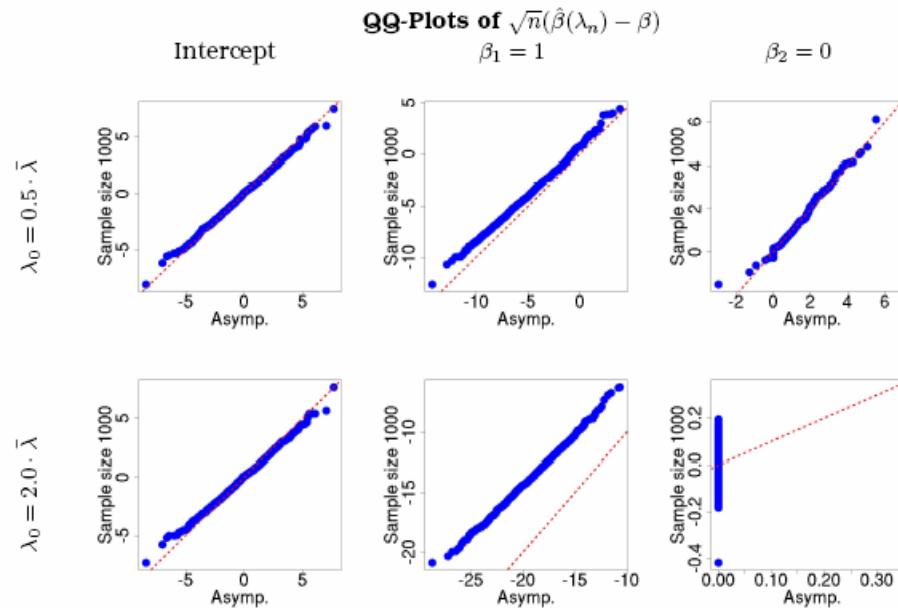


Figure 1: QQ-Plots of $\sqrt{n}(\hat{\beta}(\lambda_n) - \beta)$: Each panel contains a QQ-plot of $\sqrt{n}(\hat{\beta}(\lambda_n) - \beta)$ vs the minimizer of $V(\mathbf{u})$ for each of the coefficients (and intercept) and a specific point in the path based on 1,000 runs of each. The first, second and third columns contain the results for the intercept (not-penalized), the non-zero coefficient ($\beta_1 = 1$) and the zero coefficient ($\beta_2 = 0$). Each row contains the result for a specific value of the regularization parameter λ .

Penalty convergence (M-Estimator)

The first block $\beta_{\mathcal{I}_1} \in \mathbb{R}^q$ consists of the non-zero terms in β , while the second has $\beta_{\mathcal{I}_2} = \mathbf{0} \in \mathbb{R}^{p-q}$:

$$\begin{aligned}\beta &= [\beta_{\mathcal{I}_1}^T, \beta_{\mathcal{I}_2}^T]^T \in \mathbb{R}^q \times \mathbb{R}^{p-q} \\ &= [\beta_{\mathcal{I}_1}^T, \mathbf{0}^T]^T\end{aligned}$$

The Hessian matrix $\mathbf{H}(\beta)$ of the risk function is partitioned accordingly as:

$$\mathbf{H}(\beta) = \begin{bmatrix} \mathbf{H}_{\mathcal{I}_1\mathcal{I}_1}(\beta) & \mathbf{H}_{\mathcal{I}_1\mathcal{I}_2}(\beta) \\ \mathbf{H}_{\mathcal{I}_1\mathcal{I}_2}(\beta)' & \mathbf{H}_{\mathcal{I}_2\mathcal{I}_2}(\beta) \end{bmatrix}$$

Theorem 10. (Model selection consistency for ℓ_1 -penalized M-estimators)

Assume the assumptions in the penalty function is the ℓ_1 -norm of β , $\lim_{n \rightarrow \infty} \frac{\lambda_n}{n} = 0$ and

$\lim_{n \rightarrow \infty} \frac{\lambda_n}{n^{\frac{c+1}{2}}} = \infty$, with $c \in (0, 1)$.

If $\|\mathbf{H}_{\mathcal{I}_2\mathcal{I}_1}(\beta)\mathbf{H}_{\mathcal{I}_1\mathcal{I}_1}(\beta)^{-1}\text{sign}(\beta_{\mathcal{I}_1})\|_{\infty} \leq 1$, then $\lim_{n \rightarrow \infty} \mathbb{P}(\text{sign}(\beta_j) = \text{sign}(\hat{\beta}_j^{(n)}(\lambda_n))) = 1$.

Sign selection

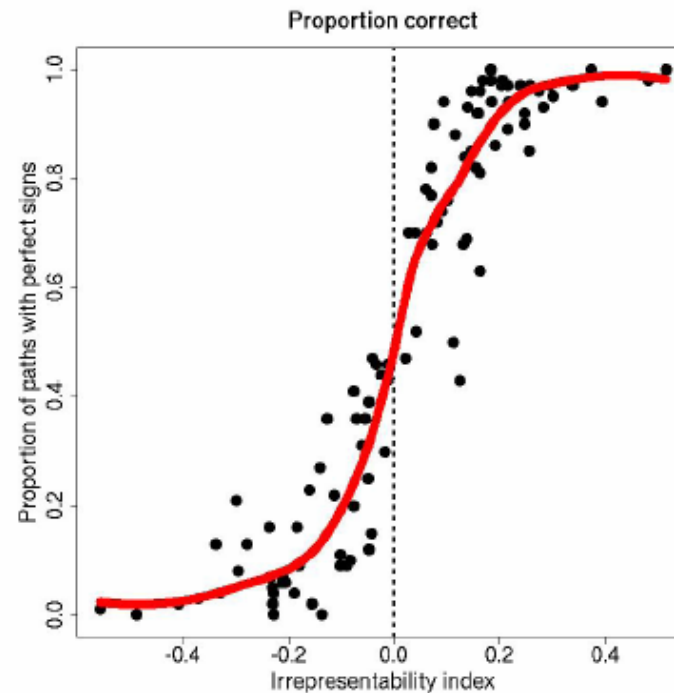


Figure 2: Correct sign selection frequency vs. Irrepresentability Index: On the figure above, each point is the the proportion of times (based on 1,000 replicates) that the ℓ_1 -penalized logistic regression path contained coefficients with all signs correct for a $p = 32$ dimensional predictor based on a sample of size $n = 10,000$. The covariance of each design was sampled from a Wishart distribution $(\mathbf{I}_{32}, 32)$. The simulation supports the result in Theorem 10 that the path contains estimates with correct signs asymptotically when the irrepresentability index $\varrho(\mathbf{H}) = 1 - \left\| \mathbf{H}_{\mathcal{I}_2, \mathcal{I}_1} \mathbf{H}_{\mathcal{I}_2, \mathcal{I}_1}^{-1} \text{sign}(\beta_{\mathcal{I}_1}) \right\|_{\infty}$ is positive.

Conclusion

- Extend Knight&Fu results to more general M-estimators and a broad family of penalty functions.
- Derive asymptotic approximations to convex twice continuously differentiable loss function.
- Establish a generalized “irrepresentable condition” on the Hessian of the risk function for sign model selection consistency of a broader family of loss functions penalized by the l_1 -norm in the parametric case.

Future work

- Extend our insights for more penalized functions.
- Extend our asymptotic work to non parametric ($p \rightarrow \infty$) case.

Thanks

wangxingscy@gmail.com

中国人民大学统计学院

School Of Statistics, Renmin University Of China