Penalized Generalized Estimating Equations for Variable Selection with Longitudinal Data

John Dziak and Runze Li
jjd264@psu.edu

http://www.stat.psu.edu/∼jdziak

Methodology Center, Pennsylvania State University

Overview of Talk

- Review of some common penalty functions
- Penalized GEE
- Theoretical results
- Empirical Results

Classic Penalized Least Squares

- We often do regression by minimizing some $\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})$ such as RSS; equiv. to $-2\ell(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y})$
- To introduce shrinkage and/or selection we minimize $\mathbb{D}$ plus a penalty
- Information criteria: $\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \lambda df$
- Ridge: $\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$

LASSO (Tibshirani 1996)

- Absolute value penalty – Minimize

$$\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \mathcal{P}(\boldsymbol{\beta}, \lambda)$$

$$\text{with } \mathcal{P}(\boldsymbol{\beta}, \lambda) = \lambda \sum_{j=1}^{d} |\beta_j|$$

- in between a count (subset selection – sparsity) and a sum of squares (ridge regression – regularization and stability)
- $\therefore$ combines shrinkage and deletion – also combines selection and estimation

SCAD (Fan and Li 2001)

Minimize

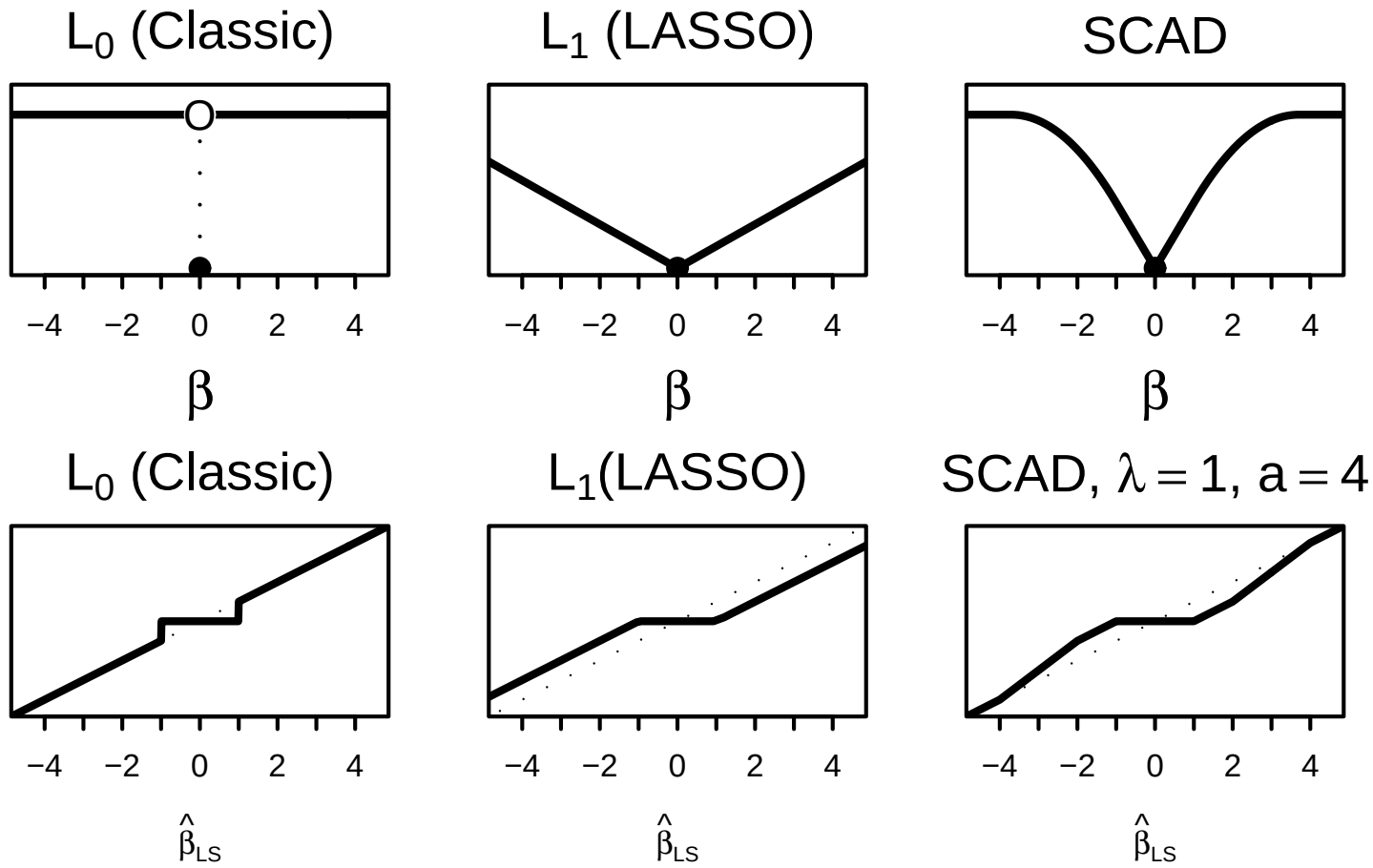
$$\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + n\mathcal{P}(\boldsymbol{\beta}, \lambda)$$

with $\mathcal{P}(\boldsymbol{\beta}, \lambda) = \sum_{j=1}^{d} p_j(|\beta_j|)$ such that

$$p_j\left(|\beta_j|\right) = \begin{cases} \lambda|\beta| & \text{if } 0 \leq |\beta| < \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta|-a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\beta| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| \geq a\lambda \end{cases}$$

SCAD (Fan and Li 2001)

Like LASSO, but with lower bias (because penalty is bounded) $\rightarrow$ "Oracle property"



SCAD (Fan and Li 2001)

- Other penalties: Garotte (Breiman 1995), bridge regression (see Fu 1998), adaptive Lasso (Zou 2006)

SCAD (Fan and Li 2001)

- Other penalties: Garotte (Breiman 1995), bridge regression (see Fu 1998), adaptive Lasso (Zou 2006)
- Choose λ by optimizing a tuning function (e.g., GCV or pseudo-BIC)

Need for Extensions to GEE

- What if a fully parametric likelihood function cannot be specified? Specifically, consider longitudinal data with form of the correlation structure only provisionally specified.

Need for Extensions to GEE

- What if a fully parametric likelihood function cannot be specified? Specifically, consider longitudinal data with form of the correlation structure only provisionally specified.
- How can we do estimation and variable selection without a likelihood (and hence without a penalized likelihood)?

GEE (Liang and Zeger 1986)

- We can use some variation on penalized GEE:
  $\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1/2} \mathbf{R}_i^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) + N\dot{\mathcal{P}}(\boldsymbol{\beta}) = \mathbf{0}$
  where $\mathbf{A}$ is the diagonal matrix of marginal variances, $\mathbf{D} = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta} = \mathbf{AX}$, and $\mathbf{R}$ is a structured working covariance matrix
- We could think of minimizing

$$\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + N\mathcal{P}(\boldsymbol{\beta}, \lambda)$$

  with
  $\mathbb{D} = \sum_{i=1}^{n} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T \mathbf{A}_i^{-1/2} \mathbf{R}_i^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$

GEE (Liang and Zeger 1986)

- GEE analogs have recently been developed for AIC (Pan 2001, *Biometrics*), C_p (Cantoni et al., 2005, *Biometrics*), and LASSO (Fu 2003, *Biometrics*)

GEE (Liang and Zeger 1986)

- GEE analogs have recently been developed for AIC (Pan 2001, *Biometrics*), C_p (Cantoni et al., 2005, *Biometrics*), and LASSO (Fu 2003, *Biometrics*)
- Possibilities for BIC (see Pauler 1998, *Biometrika*; Jiang and Liu 2004, *Journal of Statistical Planning and Inference*) – but complication of $\log(n)$ versus $\log(N)$

GEE (Liang and Zeger 1986)

- GEE analogs have recently been developed for AIC (Pan 2001, *Biometrics*), C_p (Cantoni et al., 2005, *Biometrics*), and LASSO (Fu 2003, *Biometrics*)
- Possibilities for BIC (see Pauler 1998, *Biometrika*; Jiang and Liu 2004, *Journal of Statistical Planning and Inference*) – but complication of $\log(n)$ versus $\log(N)$
- We compared these and SCAD

Thm. 3.1: $\sqrt{n}$-consistency for PGEE

For LASSO with $\lambda = \sqrt{n}$ and SCAD with $\lambda = o_p(1)$ (or other penalties; see Chapter 3 for geometric requirements on penalties), there exists a solution s.t.

$$\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right\| = O_p(n^{-1/2})$$

Theorem 3.1: $\sqrt{n}$-consistency

Proof is by showing a local minimum of $\mathbb{D}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + N\mathcal{P}(\boldsymbol{\beta}, \lambda)$ exists within a shrinking ball around the true parameter, resting on assumption of positive definite $\mathbf{K} = \mathrm{E}\mathbf{K}_n = \mathrm{E}\left(\frac{1}{n}\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)$. Note that Theorem 3.1 directly implies that $w.p. \rightarrow 1$, all significant coefficients will be included in the model (see Lemma 3.1).

Lemma 3.2: Sparsity for PGEE

With SCAD and $\lambda \to 0$, $\sqrt{n}\lambda \to \infty$, then there exists a $\sqrt{n}$-consistent sequence of estimators $\hat{\boldsymbol{\beta}}$ such that, $w.p. \to 1$, all truly zero coefficients are estimated at zero.

Thus, SCAD with $\lambda \to 0, \sqrt{n}\lambda \to \infty$ is consistent like BIC.

Proof is by showing that when $\beta_j = 0$ and $\hat{\beta}_j \neq 0$, the derivative $\frac{\partial}{\partial \beta_j}\mathbb{D} + N\frac{\partial}{\partial \beta_j}\mathcal{P}$ has reverse sign of $\hat{\boldsymbol{\beta}}$, i.e., the contribution to the penalized loss function is larger when an estimate is very near zero than when it is zero.

Thm. 3.2: Asymptotic Normality

Let active and inactive coefficients be $\boldsymbol{\beta}_{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{N}}$.
Solution to
$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1/2} \mathbf{R}_i^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) + N\dot{\mathcal{P}}(\boldsymbol{\beta}) = \mathbf{0}$
has $\boldsymbol{\beta}_{\mathcal{N}} = \mathbf{0}$, $wp \to 1$, and has
$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \hat{\boldsymbol{\beta}}_{0\mathcal{A}} + o_p(1)\right) \xrightarrow{L} \mathrm{N}(\mathbf{0}, \boldsymbol{\Phi})$...

Thm. 3.2: Asymptotic Normality

where

$$\mathbf{\Phi} = \left(\mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}^T + \frac{N}{n} \mathcal{P}''(\boldsymbol{\beta}_0) \right)^{-1} \ldots$$

$$\left(\mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{C}_n \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}^T \right) \left(\mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}^T + \frac{N}{n} \mathcal{P}''(\boldsymbol{\beta}_0) \right)$$

with $\mathbf{K}_n = \frac{1}{n} \sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$, and $\mathbf{D}_i = \mathbf{A}_i \mathbf{X}_i$, $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i$ and subscript $\mathcal{A}$ indicating active (nonzero) coefficients.

... Oracle Property

- If the response covariance is correctly specified then $\mathbf{\Phi} \approx \mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}^T$, e.g., for linear models this is $\approx \sigma^2 \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T = \sigma^2 \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$

... Oracle Property

- If the response covariance is correctly specified then $\mathbf{\Phi} \approx \mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}^T$, e.g., for linear models this is $\approx \sigma^2 \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}^T = \sigma^2 \mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}}$
- Sparsity + Sensitivity + Asymptotic Normality + Low Bias = “Oracle Property”

Gaussian Simulation

- 200 simulations
- $\mathbf{y}_i = 20 + \mathbf{X}_i\boldsymbol{\beta} + \sigma\epsilon_i$, $\sigma = 3$
- 50 subjects with 7 observations each
- $\boldsymbol{\beta} = [3, 1, 0, 0, 2, 0, 0, 0, .5, 0]$
- ϵ_i AR(1)

Results

Using working independence...				
	Cor	Ove	Und	Mis
Pan AIC	0.33	0.27	0.20	0.21
Cantoni C_p	0.35	0.26	0.21	0.19
Naïve BIC(n)	0.36	0.13	0.35	0.17
Naïve BIC(N)	0.35	0.03	0.50	0.13
LASSO, BIC(N)	0.08	0.85	0.00	0.08
SCAD, BIC(N)	0.35	0.21	0.33	0.12
With the correct working structure...				
	Cor	Ove	Und	Mis
Pan AIC	0.29	0.46	0.07	0.19
Cantoni C_p	0.33	0.41	0.08	0.19
Naïve BIC(n)	0.45	0.04	0.41	0.11
Naïve BIC(N)	0.33	0.00	0.63	0.05
LASSO, BIC(N)	0.01	0.99	0.00	0.01
SCAD, BIC(N)	0.40	0.51	0.05	0.05

Binary simulation

- 100 simulations
- 150 subjects with 10 observations each
- Used bindata package to generate exchangeably correlated binary data
- $P(Y_{ij} = 1) = g^{-1}(-2 + \mathbf{X}_{ij}\boldsymbol{\beta})$
- $\boldsymbol{\beta} = [1, \frac{1}{3}, 0, 0, \frac{2}{3}, 0, 0, 0, \frac{1}{6}, 0]^T$

Results

	Mean MSE			Correct Subsets		
	Ind	Ar	CS	Ind	Ar	CS
Full model	0.97	0.74	0.70	0.00	0.00	0.00
Pan AIC	0.79	0.58	0.57	0.29	0.26	0.30
Light BIC	0.68	0.55	0.60	0.49	0.51	0.50
Heavy BIC	0.73	0.64	0.67	0.35	0.34	0.37
LASSO, heavy BIC	0.75	0.70	0.73	0.13	0.10	0.18
SCAD, heavy BIC	0.71	0.60	0.62	0.34	0.35	0.40

Review of Simulations

Simulations show penalized GEE can be used effectively with Gaussian or binary data and a variety of correlation structures. SCAD works somewhat similarly to BIC in the GEE case; LASSO is not very parsimonious.

Discussion

- Several options exist for variable selection in GEE models
- There are many options for implementation – still working on this
- Remaining issues include robustness, covariance selection, and error estimation and model uncertainty

Important References

·Cantoni, Flemming, and Ronchetti (2005). Variable selection for marginal longitudinal generalized linear models. **Biometrics**.

·Fan and Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. **JASA**.

·Fu (2003). Penalized estimating equations. **Biometrics**.

·Gurka (2006). Selecting the best linear mixed model under REML. **Am. Stat.**

·Jiang and Liu (2004). Consistent model selection based on parameter estimates. **J. Stat. Planning and Inference**.

·Pan (2001) Akaike's Information Criterion in generalized estimating equations. **Biometrics**.

·Pauler (1998). The Schwarz criterion and related methods for normal linear models. **Biometrika**.

·Tibshirani (1996). Regression shrinkage and selection via the Lasso. **JRSS-B**.