L^γ Penalty Models

Computation And Applications

Part III

Wenjiang Fu

wfu@stat.tamu.edu

http://stat.tamu.edu/∼wfu.

Department of Statistics, Texas A&M University

EXTENSION TO NON–GAUSSIAN DATA

Biomedical research: responses are often non-Gaussian

- ◇ Binary: disease (1) v.s. normal (0), death (1) vs.. live (0).
- ◇ Counts: number of traffic accidents on one piece of high way, etc.
- ◇ Ordinal: grade of pain, etc.
- ◇ Categorical: hobby, music, sports, etc.

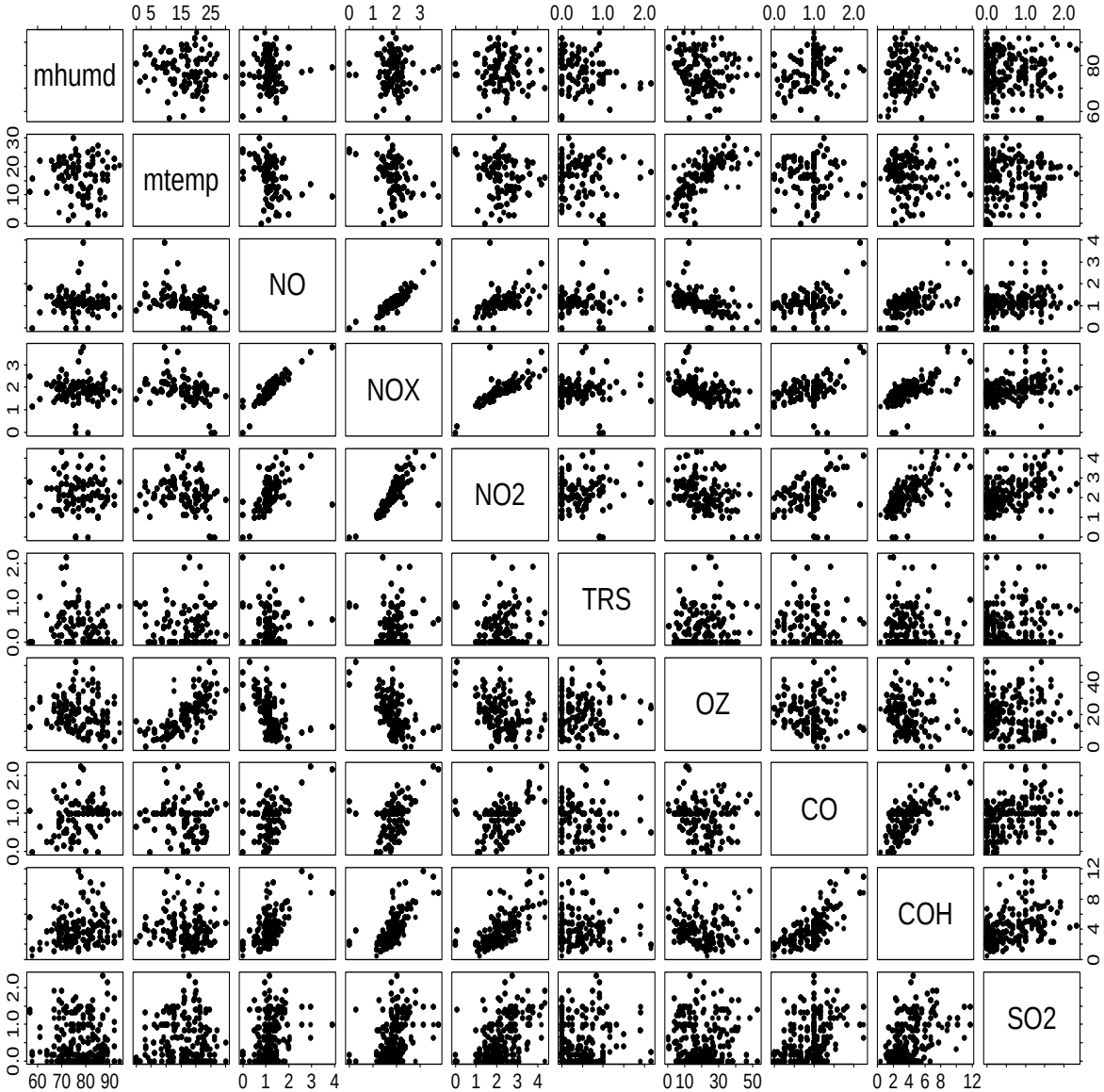
Biomedical studies collect lots of data,

including demographic, socioeconomic, health related data and family history. Variables tend to be collinear.

Need to extend L^γ penalty model from linear to generalized linear models, and correlated observation models in longitudinal studies.

EXTENSION TO NON–GAUSSIAN DATA

An example of air pollution impact on asthma patients



EXTENSION TO NON–GAUSSIAN DATA

Generalized linear models (McCullagh and Nelder 1991)

$\boldsymbol{Y} \sim$ distribution $\boldsymbol{f}$ in exponential family with loglik:

$$\boldsymbol{l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)}$$

GLM: $\boldsymbol{g(\theta) = \eta;}$ $\boldsymbol{\eta = X\beta}$.

Q: How to apply penalty models to GLM?

To replace RSS in linear model with model deviance:

$\boldsymbol{D(\theta; y) = 2l(y; y) - 2l(\mu; y)}$.

Penalized GLM:

$$\min_{\boldsymbol{\beta}} \left\{ \boldsymbol{Dev(\theta(\beta); y) + \lambda \sum |\beta_j|^{\gamma}} \right\}$$

For given $\boldsymbol{\lambda \geq 0}$, $\boldsymbol{\gamma \geq 1}$, there exists a unique estimator. $\boldsymbol{\widehat{\beta}_{\mathbf{rdg}}}$, $\boldsymbol{\widehat{\beta}_{\mathbf{lss}}}$ and $\boldsymbol{\widehat{\beta}_{\mathbf{brdg}}}$ exist.

EXTENSION TO NON–GAUSSIAN DATA

Computation for penalty estimators

Incorporate the penalty algorithms (M–N–R or shooting) in the IRLS (Iteratively reweighted least squares) procedure.

IRLS (McCullagh and Nelder 1991):

- ◇ Let $\widehat{\mu}, \widehat{\eta}$ be current estimate.
- ◇ Form adjusted dependent variable
  $z = \widehat{\eta} + (y - \widehat{\mu})\,(d\eta/d\mu)$
  and weights $W^{-1} = (d\eta/d\mu)^2\, V$.
- ◇ Regress z on X with weight W or
  regress $W^{1/2}z$ on $W^{1/2}X$ to yield new estimator $\widehat{\beta}$
  and a new linear predictor η.
- ◇ Repeat the above until convergence.

Select tuning parameter via NLGCV

Modify NLGCV by replacing RSS in (6) with model deviance:

$$\mathbf{NLGCV} = \frac{\boldsymbol{Dev}(\boldsymbol{\mu}, \boldsymbol{y})}{\boldsymbol{n}(1 - \boldsymbol{ps}/\boldsymbol{n})^{\mathbf{2}}}$$

where s is the standard shrinkage rate defined as before.

$$\boldsymbol{s} = \frac{\|\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\gamma})\|_{\gamma}}{\|\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}\|_{\gamma}}$$

where $\widehat{\beta}^{(0)}$is GLM estimator with no-penalty, and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ the estimator with penalty.

EXTENSION TO GEE MODEL

Longitudinal studies:

K subjects, each has multiple obs.

subject	time		
k	t_1	$\cdots$	t_n
1	(x_{11}, y_{11})		(x_{1t_1}, y_{1t_1})
		$\cdots$	
K	(x_{K1}, y_{K1})		(x_{Kt_K}, y_{Kt_K})

Observations within each subject are correlated.

Purpose: to study how response Y depends on variables x.

Generalized estimating equations (GEE) model (Liang and Zeger 1986, Zeger and Liang 1986).

Idea: to incorporate a working correlation structure between observations into estimating equations.

EXTENSION TO GEE MODEL

GEE model – independent observations:

When observations are independent and have distribution in the exponential family, the estimating equations are

$$S(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T v_i^{-1} \left[y_i - \mu_i(\beta) \right]$$

where $v_i = \mathbf{var}(Y_i)$.

Correlated observations – to specify

a working correlation structure to replace v_i.

EXTENSION TO GEE MODEL

GEE model:

Based on marginal distributions of response

$f(y_{kt}) = \exp[\{y_{kt}\theta_{kt} - a(\theta_{kt}) + b(y_{kt})\}\phi]$,

GEE: $\sum_{k=1}^{K} D_k^T V_k^{-1} S_k = 0$

with working covariance matrix: $V_k = A_k^{1/2} R(\alpha) A_k^{1/2}/\phi$,
$D_k = d\{a_k'(\theta)\}/d\beta = A_k \Delta_k X_k$, $\Delta_k = \text{diag}(d\theta_{kt}/d\eta_{kt})$,
and $S_k = y_k - a_k'(\theta)$.

EXTENSION TO GEE MODEL

Advantages of the GEE model:

1). Based on marginal likelihood, requires no joint likelihood.

2). Estimation is consistent even with incorrect correlation structure specified.

$$\sqrt{K}(\widehat{\beta} - \beta) \rightarrow_d N(0, V).$$

V is a var/cov matrix and is estimated with a sandwich estimator in (Liang and Zeger 1986).

3). Correct specification of correlation structure increases efficiency.

EXTENSION TO GEE MODEL

Data analysis of asthma study via GEE model :

39 asthmatics observed on 21 consecutive days .

Estimates and SE of major pollutants via GEE model.

Covar	Est(se)	Covar	Est(se)
intercept	-2.659(.464)		
mhumd	0.044(.062)	TRS	-0.082(.116)
mtemp	-0.438(.374)	OZ	-0.269(.174)
NO	-1.015(.405)	CO	0.591(.155)
NO2	-0.787(.175)	COH	-0.106(.426)
NOX	0.975(.464)	SO2	0.413(.106)

Q: What's wrong?

EXTENSION TO GEE MODEL

Collinearity in longitudinal studies :

Nothing wrong with the GEE model. But strong collinearity!

Q: How to apply penalty to GEE model?

Recall: 1). GEE requires no joint likelihood;

2). Penalty model: $\min\{Dev + \sum |\beta_j|^\gamma\}$.

Deviance comes from joint likelihood. So need to extend penalty model from joint likelihood–dependent to joint likelihood–independent.

Theorem 1 provides the theoretical support, which only requires a Jacobian condition.

EXTENSION TO GEE MODEL

Extension of penalty model:

$$\begin{cases} \boldsymbol{F}_1(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{\lambda} \boldsymbol{d}(\boldsymbol{\beta}_1, \boldsymbol{\gamma}) &= \mathbf{0} \\ \cdots & \\ \boldsymbol{F}_{\boldsymbol{p}}(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{y}) + \boldsymbol{\lambda} \boldsymbol{d}(\boldsymbol{\beta}_{\boldsymbol{p}}, \boldsymbol{\gamma}) &= \mathbf{0}, \end{cases} \tag{7}$$

where $\boldsymbol{d}(\boldsymbol{\beta}_j, \boldsymbol{\gamma}) = \boldsymbol{\gamma} |\boldsymbol{\beta}_j|^{\boldsymbol{\gamma}-\mathbf{1}} \mathbf{sign}(\boldsymbol{\beta}_j)$.

Theorem 5 If $\boldsymbol{F} = (\boldsymbol{F}_1, \ldots, \boldsymbol{F}_p)$ are continuously differentiable, and Jacobian matrix $\boldsymbol{\partial F / \partial \beta}$ is positive-semi-definite. Then for given $\boldsymbol{\lambda} > \mathbf{0}, \boldsymbol{\gamma} > \mathbf{1}$, there exists a unique solution $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ of equations (7). $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\gamma})$ is continuous and the limit $\mathbf{lim}_{\boldsymbol{\gamma} \to \mathbf{1}+} \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \mathbf{1}+)$ exists.

EXTENSION TO GEE MODEL

Penalized estimating equations:

Definition

Equation system (7) is called penalized estimating equations. The solution $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, \gamma)$ is said to be the bridge estimator with $\gamma > 1$, and the limit $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}, 1+)$ is said to be the lasso estimator.

This new definition is independent of joint likelihood and thus can be applied to the GEE models without difficulty. Only need to verify the Jacobian condition.

EXTENSION TO GEE MODEL

GEE satisfies Jacobian condition:

Let $\boldsymbol{H} = \partial(-\sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{S}_k)/\partial\boldsymbol{\beta}$. Since

$$\begin{aligned}\partial \boldsymbol{S}_k/\partial\boldsymbol{\beta} &= (\partial \boldsymbol{S}_k/\partial\boldsymbol{\theta}_k)(\partial\boldsymbol{\theta}_k/\partial\boldsymbol{\eta}_k)(\partial\boldsymbol{\eta}_k/\partial\boldsymbol{\beta}) = -\boldsymbol{A}_k\boldsymbol{\Delta}_k\boldsymbol{X}_k \\ &= -\boldsymbol{D}_k,\end{aligned}$$

$$\frac{\boldsymbol{H}}{\boldsymbol{K}} = -\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\partial \boldsymbol{D}_k^T}{\partial\boldsymbol{\beta}}\boldsymbol{V}_k^{-1} + \boldsymbol{D}_k^T\frac{\partial \boldsymbol{V}_k^{-1}}{\partial\boldsymbol{\beta}}\right)\boldsymbol{S}_k + \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\boldsymbol{D}_k$$

The second term is positive-definite and converges to psd.

By regularity conditions (Liang and Zeger, 1986), $(\partial \boldsymbol{D}_k^T/\partial\boldsymbol{\beta})\boldsymbol{V}_k^{-1} + \boldsymbol{D}_k^T(\partial \boldsymbol{V}_k^{-1}/\partial\boldsymbol{\beta})$ is bounded.

$\boldsymbol{S}_k$ are independent with $\mathrm{E}(\boldsymbol{S}_k) = \boldsymbol{0}$ and finite variance $\mathrm{Var}(\boldsymbol{S}_k) \leq \boldsymbol{C} < \infty$ indep. of $\boldsymbol{k}$. The first term converges to $\boldsymbol{0}$ in $\boldsymbol{L^2}$ and in probability by Weak LLN (Durrett, 1991, p. 29).

EXTENSION TO GEE MODEL

Asymptotics of PENEE estimators :

Regularity conditions: $\partial^2 F/\partial\beta^2$ exists; Jacobian $\partial F/\partial\beta$ pos.–def.; $F = \sum_{k=1}^{K} X_k^T G_k(\beta; X, y)$ (Yuan and Jennrich, 2000), G_k i.i.d. vectors of r.fun. w. finite mean $G_0(\beta)$. Matrix X_k is bounded for $k \geq 1$; limit $\lim_{K\to\infty} \sum_{k=1}^{K} X_k/K = X_0$ exists. $X_k X_k^T$ and the limit $\lim_{K\to\infty} \sum_{k=1}^{K} X_k X_k^T/K$ are nondegenerate.

Theorem 6 Assume $\lambda_K = o(\sqrt{K})$. The PENEE (7) for fixed $\gamma \geq 1$ yields $\widehat{\beta}$ with

$$\sqrt{K}\left(\widehat{\beta} - \beta_\infty\right) \longrightarrow_d N(0, \Sigma) \qquad as \quad K \to \infty,$$

where β_∞ is true parameter of (7) and Σ is a pos.–def. var/cov matrix, see van der Vaart (1998, pp. 51–52) for details of Σ.

EXTENSION TO GEE MODEL

Computations for PENEE model :

Recall: GEE takes IRLS procedure to fit the models (Liang and Zeger 1986).

Incorporate the penalization procedure (M–N–R or shooting) into the IRLS procedure to obtain bridge estimators for PENEE models with fixed $\lambda > 0$ and $\gamma \geq 1$.

Selection of tuning parameter λ

Since NLGCV (6) depends on model deviance, which does not exist in GEE models, need to modify NLGCV.

Idea: to incorporate the working correlation structure into deviance to make it into a weighted deviance.

EXTENSION TO GEE MODEL

Motivation for Weighted deviance :

Assume $\boldsymbol{Y}$ are correlated responses from model $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ with $\varepsilon \sim \boldsymbol{N}(\boldsymbol{0}, \Sigma)$, where Σ is a non-diagonal variance-covariance matrix. To apply GCV for indep. responses, take a transformation $\boldsymbol{Z} = \boldsymbol{PY}$, where $\boldsymbol{P} = \Lambda^{-1/2}\boldsymbol{Q}$ satisfying $\boldsymbol{Q}\Sigma\boldsymbol{Q}^T = \Lambda$, a diagonal matrix. Then $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{PX}\beta, \boldsymbol{I})$. Apply GCV to $\boldsymbol{Z}$,

$$\begin{aligned}\mathrm{RSS} &= (\boldsymbol{Z}-\boldsymbol{PX}\beta)^{\boldsymbol{T}}(\boldsymbol{Z}-\boldsymbol{PX}\beta) = (\boldsymbol{Y}-\boldsymbol{X}\beta)^{\boldsymbol{T}}\boldsymbol{P}^{\boldsymbol{T}}\boldsymbol{P}(\boldsymbol{Y}-\boldsymbol{X}\beta) \\ &= (\boldsymbol{Y}-\boldsymbol{X}\beta)^{\boldsymbol{T}}\Sigma^{-1}(\boldsymbol{Y}-\boldsymbol{X}\beta).\end{aligned}$$

Thus, GCV can be applied to correlated observations $\boldsymbol{Y}$ by incorporating the correlation structure in the residuals.

EXTENSION TO GEE MODEL

Weighted deviance :

Recall: For GLMs, $Dev = \sum r_i^2$, r_is are deviance residuals. Although Dev does not exist in GEE models, deviance residuals r_{kt} can be computed using marginal likelihood.
$r_{kt} = \text{sign}(y_{kt} - \widehat{\mu}_{kt})\sqrt{-2LogL(y_{kt}, \widehat{\mu}_{kt})}$.
Define weighted deviance

$$\mathbf{WDev}(\boldsymbol{\lambda}, \boldsymbol{\gamma}) = \sum_{k=1}^{K} {r_k}^T R_k^{-1} r_k,$$

where r_k is deviance residual vector of subject k.
The weighted deviance reduces to deviance with independent observations.

EXTENSION TO GEE MODEL

Effective number of observations :

Within subject observations are correlated, and the number of observations effective in the model needs to be adjusted.

A simple adjustment method:

$$N = \sum_{k=1}^{K} \frac{n_k^2}{|R_k|},$$

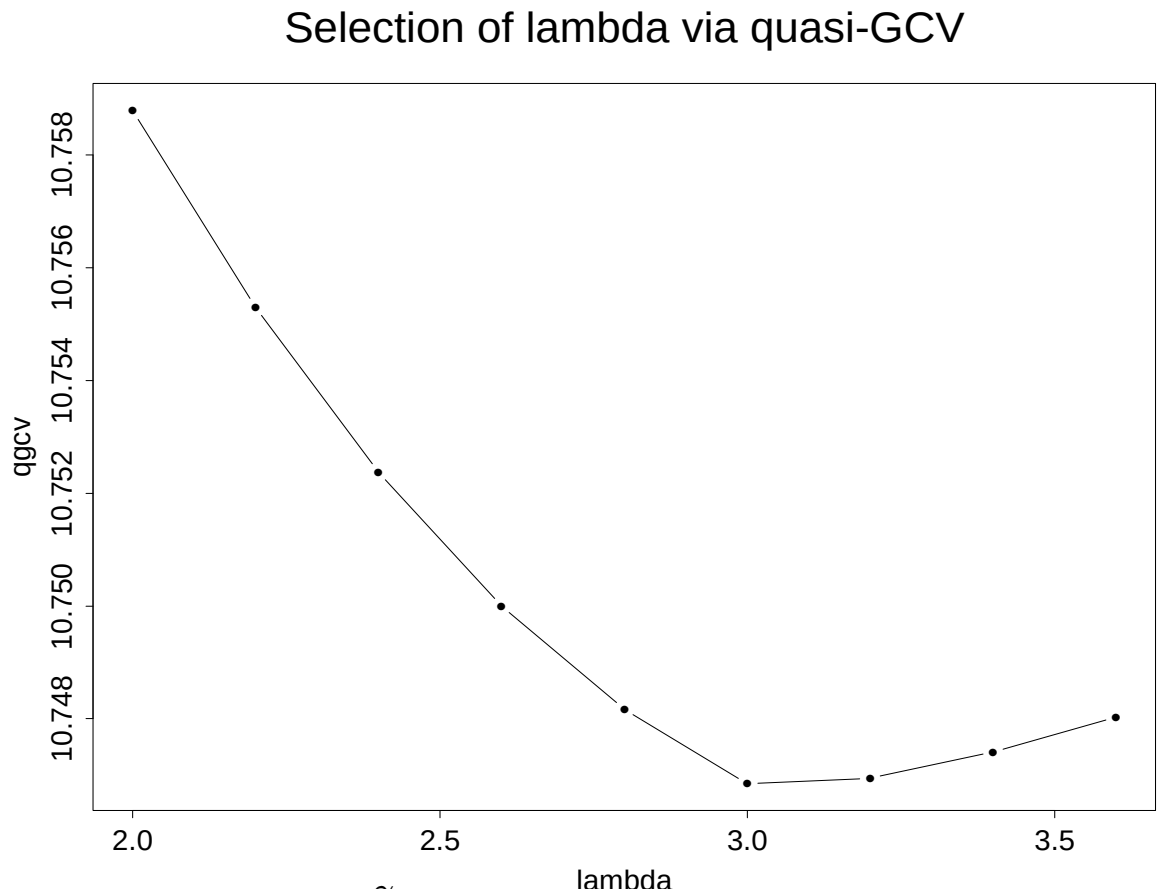
where n_k is the number of observations of subject k, $|R_k| = \sum \rho_{ij}$ of correlation matrix $R_k = (\rho_{ij})$ for subject k. Thus N is usually between K and the total number of observations depending on the correlation.

EXTENSION TO GEE MODEL

Quasi–GCV :

$$\mathbf{QGCV} = \frac{\boldsymbol{WDev}(\boldsymbol{\lambda}, \boldsymbol{\gamma})}{\boldsymbol{K}(1 - \boldsymbol{ps}/\boldsymbol{N})^{\mathbf{2}}}$$

Select λ for fixed γ. For the same reason, QGCV cannot be used for the selection of γ.



EXTENSION TO GEE MODEL

MSEs from 1000 simulation runs

$\boldsymbol{p = 5}$ covariates, $\boldsymbol{K = 10}$ subjects, $\boldsymbol{t = 5}$ obs each.

Model	No-penalty GEE	Lasso GEE	Ridge GEE
Parameter*	$\lambda = 0$	$\gamma = 1$	$\gamma = 2$
β_1	.0264(.0018)	.0444(.0032)	.0126(.0009)
β_2	.0069(.0006)	.0052(.0005)	.0039(.0003)
β_3	.0065(.0005)	.0042(.0004)	.0036(.0002)

* $\boldsymbol{\beta_1 = (1, 0.5, -0.2, 1, -1)}$, $\boldsymbol{\beta_2 = (1, -0.5, 0, 0, 0)}$ and $\boldsymbol{\beta_3 = (1, 0, 0, 0, 0)}$.

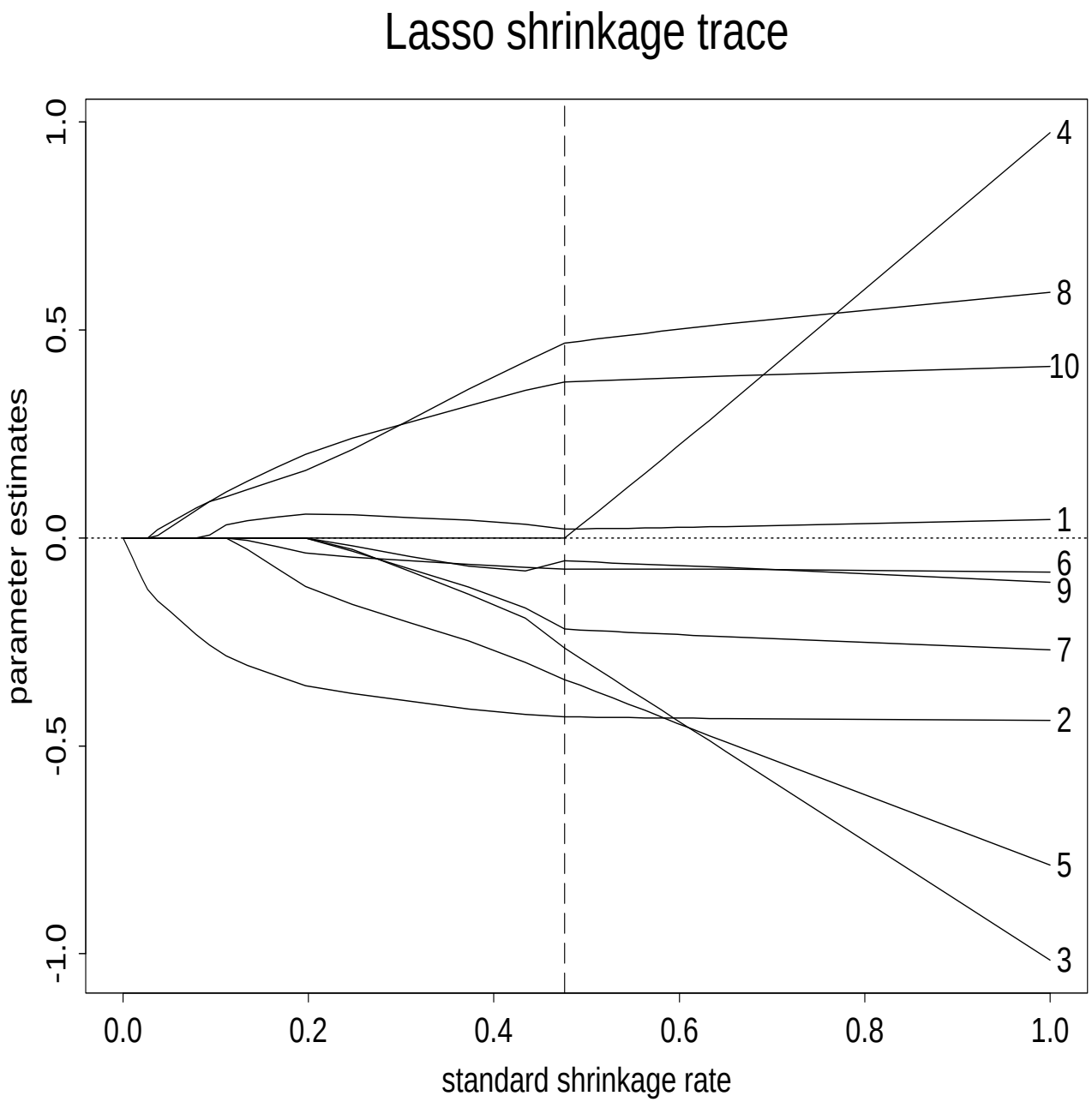
Ten sets of Poisson responses generated for each highly correlated regression matrix $\boldsymbol{X}$ generated with random numbers.
$\mathbf{MSE} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\boldsymbol{T}} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Repeat 100 times.

EXTENSION TO GEE MODEL

Air pollution data analysis

	No penalty $\lambda = 0$	Lasso penalty $\lambda = 3.0$	Ridge penalty $\lambda = 3.2$
Intercept	-2.659(.464)	-2.634(.644)	-2.640(.658)
meanhumd	0.044(.062)	0.021(.077)	0.052(.081)
meantemp	-0.438(.374)	-0.430(.567)	-0.428(.499)
NO	-1.015(.405)	-0.265(.308)	-0.335(.213)
NO2	-0.787(.175)	-0.340(.393)	-0.373(.215)
NOX	0.975(.464)	0(.094)	0.159(.200)
TRS	-0.082(.116)	-0.075(.166)	-0.087(.159)
OZ	-0.269(.174)	-0.219(.236)	-0.201(.155)
CO	0.591(.155)	0.468(.190)	0.489(.160)
COH	-0.106(.426)	-0.056(.513)	-0.139(.419)
SO2	0.413(.106)	0.374(.127)	0.381(.113)

EXTENSION TO GEE MODEL



REFERENCES

Durrett, R. (1991). *Probability Theory and Examples*. Belmont: Wadsworth.

Fu, W.J. (1998) Penalized regressions: the Bridge versus the Lasso, *J. Comp. Grap. Statist.* 7: 397-416.

Fu, W.J. (2003) Penalized estimating equations, *Biometrics*, 59:126-132.

Fu, W.J. (2005) Nonlinear GCV and quasi–GCV for shrinkage models, *Statist. Plann. Infer.*, in press.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. London: Chapman and Hall.

Yuan, K. H. and Jennrich, R. I. (2000). Estimating equations with nuisance parameters: theory and applications. *Annals of Institute of Statistical Mathematics* **52**, 343-350.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.