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
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 highway, 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 $L^\gamma$ 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)*

\( Y \sim \text{distribution } f \) in exponential family with loglik:

\[
l(\theta; y) = \frac{y^{\theta} - b(\theta)}{a(\phi)} + c(y, \phi)
\]

**GLM:**  \( g(\theta) = \eta; \quad \eta = X\beta \).

Q: How to apply penalty models to GLM?

To replace RSS in linear model with model deviance:

\[
D(\theta; y) = 2l(y; y) - 2l(\mu; y).
\]

**Penalized GLM:**

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

For given \( \lambda \geq 0, \gamma \geq 1 \), there exists a unique estimator. \( \hat{\beta}_{rdg}, \hat{\beta}_{lss} \) and \( \hat{\beta}_{brdg} \) 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 $\hat{\mu}, \hat{\eta}$ be current estimate.
- Form adjusted dependent variable $z = \hat{\eta} + (y - \hat{\mu}) \frac{d\eta}{d\mu}$ and weights $W^{-1} = \frac{1}{(d\eta/d\mu)^2} V$.
- Regress $z$ on $X$ with weight $W$ or regress $W^{1/2}z$ on $W^{1/2}X$ to yield new estimator $\hat{\beta}$ and a new linear predictor $\eta$.
- Repeat the above until convergence.
EXTENSION TO NON-GAUSSIAN DATA

Select tuning parameter via NLGCV

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

\[
NLGCV = \frac{\text{Dev}(\mu, y)}{n(1 - ps/n)^2}
\]

where \( s \) is the standard shrinkage rate defined as before.

\[
s = \frac{||\hat{\beta}(\lambda, \gamma)||_\gamma}{||\hat{\beta}(0)||_\gamma}
\]

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

**Extension to GEE model**

**Longitudinal studies:**

$K$ subjects, each has multiple obs.

<table>
<thead>
<tr>
<th>subject</th>
<th>time</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$t_1$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$t_n$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$(x_{11}, y_{11})$</td>
<td>$(x_{1t_1}, y_{1t_1})$</td>
</tr>
<tr>
<td></td>
<td>$\cdots$</td>
<td></td>
</tr>
<tr>
<td>$K$</td>
<td>$(x_{K1}, y_{K1})$</td>
<td>$(x_{Kt_K}, y_{Kt_K})$</td>
</tr>
</tbody>
</table>

Observations within each subject are correlated.

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


Idea: to incorporate a working correlation structure between observations into estimating equations.
GEE model – independent observations:

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

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

where \( v_i = \text{var}(Y_i) \).

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.

\[ \sqrt{K} (\hat{\beta} - \beta) \rightarrow_d N(0, V). \]

\( V \) 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.
Data analysis of asthma study via GEE model:

39 asthmatics observed on 21 consecutive days.

Estimates and SE of major pollutants via GEE model:

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

Q: What’s wrong?
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: \( \min \{ \text{Dev} + \sum |\beta_j|^{\gamma} \} \).

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 of penalty model:

\[
\begin{align*}
F_1(\beta, X, y) + \lambda d(\beta_1, \gamma) &= 0 \\
\ldots \\
F_p(\beta, X, y) + \lambda d(\beta_p, \gamma) &= 0,
\end{align*}
\]

where \(d(\beta_j, \gamma) = \gamma|\beta_j|^{\gamma^{-1}}\text{sign}(\beta_j)\).

**Theorem 5**  
If \(F = (F_1, \ldots, F_p)\) are continuously differentiable, and Jacobian matrix \(\partial F / \partial \beta\) is positive-semi-definite. Then for given \(\lambda > 0, \gamma > 1\), there exists a unique solution \(\hat{\beta}(\lambda, \gamma)\) of equations (7). \(\hat{\beta}(\lambda, \gamma)\) is continuous and the limit \(\lim_{\gamma \to 1^+} \hat{\beta}(\lambda, \gamma) = \hat{\beta}(\lambda, 1^+)\) exists.
Penalized estimating equations:

Definition
Equation system (7) is called penalized estimating equations. The solution $\hat{\beta}(\lambda, \gamma)$ is said to be the bridge estimator with $\gamma > 1$, and the limit $\hat{\beta}(\lambda, 1^+)$ 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.
GEE satisfies Jacobian condition:

Let \( H = \partial(- \sum_{k=1}^{K} D_k^T V_k^{-1} S_k)/\partial \beta \). Since

\[
\partial S_k / \partial \beta = (\partial S_k / \partial \theta_k)(\partial \theta_k / \partial \eta_k)(\partial \eta_k / \partial \beta) = -A_k \Delta_k X_k
\]

\[
= -D_k,
\]

\[
\frac{H}{K} = \]

\[
-\frac{1}{K} \sum_{k=1}^{K} \left( \frac{\partial D_k^T}{\partial \beta} V_k^{-1} + D_k^T \frac{\partial V_k^{-1}}{\partial \beta} \right) S_k + \frac{1}{K} \sum_{k=1}^{K} D_k^T V_k^{-1} D_k
\]

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

By regularity conditions (Liang and Zeger, 1986),

\[
(\partial D_k^T / \partial \beta) V_k^{-1} + D_k^T (\partial V_k^{-1} / \partial \beta) \]

is bounded.

\( S_k \) are independent with \( \mathbb{E}(S_k) = 0 \) and finite variance

\( \text{Var}(S_k) \leq C < \infty \) indep. of \( k \). The first term converges to 0 in \( L^2 \) and in probability by Weak LLN (Durrett, 1991, p. 29).
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 $\hat{\beta}$ with

$$\sqrt{K} \left( \hat{\beta} - \beta_\infty \right) \xrightarrow{d} N(0, \Sigma) \quad \text{as} \quad K \to \infty,$$

where $\beta_\infty$ is true parameter of (7) and $\Sigma$ is a pos.–def. var/cov matrix, see van der Vaart (1998, pp. 51–52) for details of $\Sigma$. 

---

**Extension to GEE Model**
Computation 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 $\lambda$

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

$$RSS = (Z - PX\beta)^T (Z - PX\beta) = (Y - X\beta)^T P^T P (Y - X\beta) = (Y - X\beta)^T \Sigma^{-1} (Y - X\beta).$$

Thus, GCV can be applied to correlated observations $Y$ by incorporating the correlation structure in the residuals.
Weighted deviance:

Recall: For GLMs, $Dev = \sum r_i^2$, $r_i$s are deviance residuals. Although $Dev$ does not exist in GEE models, deviance residuals $r_{kt}$ can be computed using marginal likelihood.

$$ r_{kt} = \text{sign}(y_{kt} - \hat{\mu}_{kt}) \sqrt{-2 \text{LogL}(y_{kt}, \hat{\mu}_{kt})}. $$

Define weighted deviance

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

where $r_k$ is deviance residual vector of subject $k$.

The weighted deviance reduces to deviance with independent observations.
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.
Quasi–GCV:

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

Select $\lambda$ for fixed $\gamma$. For the same reason, QGCV cannot be used for the selection of $\gamma$. 
Extension to GEE Model

MSEs from 1000 simulation runs

\( p = 5 \) covariates, \( K = 10 \) subjects, \( t = 5 \) obs each.

<table>
<thead>
<tr>
<th>Model</th>
<th>No-penalty GEE</th>
<th>Lasso GEE</th>
<th>Ridge GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter*</td>
<td>( \lambda = 0 )</td>
<td>( \gamma = 1 )</td>
<td>( \gamma = 2 )</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>.0264(.0018)</td>
<td>.0444(.0032)</td>
<td>.0126(.0009)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>.0069(.0006)</td>
<td>.0052(.0005)</td>
<td>.0039(.0003)</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>.0065(.0005)</td>
<td>.0042(.0004)</td>
<td>.0036(.0002)</td>
</tr>
</tbody>
</table>

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

Ten sets of Poisson responses generated for each highly correlated regression matrix \( X \) generated with random numbers.

\[ \text{MSE} = (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \]. Repeat 100 times.
## Extension to GEE Model

### Air pollution data analysis

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


