Variable Selection with Penalized Generalized Estimating Equations

John Dziak and Runze Li
The Methodology Center
Penn State University
State College, PA

July 5, 2006

Technical Report 06-78
The Methodology Center, Pennsylvania State University

Abstract

For decades, much research has been devoted to developing and comparing variable selection methods, but primarily for the classical case of independent observations. Existing variable-selection methods can be adapted to cluster-correlated observations, but some adaptation is required. For example, classical model fit statistics such as AIC and BIC are undefined if the likelihood function is unknown (Pan, 2001). We review existing research on variable selection for generalized estimating equations (GEE, Liang and Zeger, 1986) and similar correlated data approaches. We also extend the results of Fan and Li (2001) on variable selection through nonconcave penalized likelihood, to a GEE setting. The asymptotic normality and efficiency of the SCAD-penalized GEE estimator is demonstrated.

1 Introduction

The "generalized estimating equations" (GEE) of Liang and Zeger (1986) is a very popular approach to regression in longitudinal data, providing a reasonable and straightforward extension of general linear models and has fairly good

statistical properties. It avoids the explicit specification of a likelihood function, instead requiring only the specification of the marginal mean function and of a tentative "working" structure for the covariance among responses. This is important because the exact distribution and correlation structure are often unknown for longitudinal data. However, it creates a difficulty for model selection; many traditional model selection criteria, such as AIC and BIC, need to be redefined because of the clustered structure of the observations and the lack of an explicit likelihood function. This report reviews and extends the small existing literature on penalty-based approaches to variable selection in GEE. First let us review some more traditional criteria.

2 Penalized Likelihood Variable Selection

Suppose that we wish to model the relationship between predictors $x_1, ..., x_d$ and scalar response y by fitting a linear model. We are not sure which subset of the d predictors to use, but want to choose a subset of $d_{in} \leq d$ of them which does an adequate job in prediction, using the data to choose both the size d_{in} and the particular set of d_{in} predictors; coefficients for all other predictors will be set to zero.

2.1 Classical Penalized Criteria

Many variable selection criteria, such as Mallows' C_p (Mallows, 1973), AIC (Akaike, 1973), and BIC (Schwarz, 1978), involve optimizing a fit criterion (a least squares or likelihood function) modified by a complexity penalty based on the number of free nonzero parameters. This expresses our intuition that a good model should fit well while using few parameters. Specifically, we may maximize a penalized log-likelihood:

$$\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) - \mathcal{P}(\boldsymbol{\beta}). \tag{1}$$

Here ℓ is the likelihood, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $\mathcal{P}(\boldsymbol{\beta})$ is a measurement of model complexity, generally a norm measuring the size of the parameter vector $\boldsymbol{\beta}$ in some sense.

A simple way to operationalize the size of $\boldsymbol{\beta}$ is in terms of the number d_{in} of free nonzero coefficients in $\boldsymbol{\beta}$. Then $\mathcal{P}(\boldsymbol{\beta}) = \lambda d_{in}$, $\lambda > 0$, and optimizing (1) involves making ℓ large while keeping $\boldsymbol{\beta}$ small. We could choose λ to express our view of the relative importance of parsimony versus good sample fit. AIC

and BIC are each equivalent to different choices of λ, i.e., to maximizing $\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) - \lambda d_{in}$. By convention this is usually rewritten as minimizing the penalized deviance

$$-2\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) + \lambda d_{in}, \tag{2}$$

with rescaled (doubled) λ.

Also, for linear models, minimizing (2) is equivalent to minimizing the penalized variance estimator

$$N \log(\hat{\sigma}^2) + 2\kappa d_{in} \tag{3}$$

if σ^2 is unknown. Last, if σ^2 is known (or can be consistently estimated from a full model) then (2) is also equivalent to a penalized least squares (PLS) criterion, i.e.,

$$\mathrm{RSS} + \sigma^2 \lambda d_{in} \tag{4}$$

where RSS is the sum of squared residuals. The properties of these criteria are explored by Foster and George (1994), Shao (1997), and Yang (2004, 2005); they are only briefly reviewed below.

As a caveat, note that d_{in} above is treated as the number of *predictors*, but AIC and BIC are actually defined in terms of the number of *parameters*; thus for strict correctness d_{in} should be incremented by two (one for the intercept β_0 and possibly one for σ^2). However, this usually does not matter, since β_0 and σ^2 are found in all models under consideration.

The AIC of Akaike (1973), corresponding to $\lambda = 2$ in (4), attempts to estimate Kullback-Leibler discrepancy, or distance of a model from the true likelihood function. This cannot be done by simply finding the model which gives the largest fitted log-likelihood value, because maximum likelihood tends to overfit (i.e., the parameters are selected using the same data on which their fit is tested, so that we can always make the model *seem* to fit better by adding more predictors). However, Akaike (1973) showed that a roughly unbiased estimator of $E_y E_x \left(\ln \mathrm{g}(\mathbf{X}, \mathbf{y} | \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y})) \right)$, which in turn is closely related to the Kullback-Leibler discrepancy, is $\ell(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) - d_{in}$ Thus, fitted models with low AIC provide a likelihood function closer (in the Kullback-Leibler sense) to the true likelihood function. A finite-sample correction to AIC to improve the asymptotic approximation for small N is proposed in Hurvich and Tsai (1989). The theoretical and practical properties of model selection by AIC are further reviewed by Shibata (1989) and Burnham and Anderson (2002, 2004).

Mallows' C_p (Mallows, 1973, see also George (2000), Hastie, Tibshirani, and Friedman (2001), or Faraway (2002)), which estimates future prediction error in linear models, is defined as

$$C_p = \mathrm{RSS}(\boldsymbol{\beta})/\sigma^2 + 2d_{in} - N)\hat{\sigma}^2 \tag{5}$$

It is asymptotically equivalent to the AIC for linear models, and heurisitcally acts similarly in adjusting for the overfitting that occurs when there are many free parameters, in an attempt to get a more realistic estimate of model performance.

This adjustment is actually slightly too weak; it does not take into account the fact that the subset chosen was not set a priori but was selected by a data-driven method which introduces additional bias (see Hastie et al. 2001, p. 204, and Miller 2002). Thus if we choose the model with the lowest C_p there will be a remaining tendency to overfit. This problem is addressed by the Bayesian Information Criterion (BIC).

In Bayesian model selection we can set a prior probability for each model $\mathcal{M}_i$, and prior distributions for the nonzero coefficients in each model. If we then assume that one and only one model, along with its associated priors, is correct, we can use Bayes' Theorem to find the posterior probability of each model given the data. The posterior probability of a model is its prior probability times the likelihood of the data under the model

$$\Pr(\mathcal{M}_i|\mathbf{x},\mathbf{y}) = \Pr(\mathcal{M}_i)\Pr(\mathbf{y}|\mathbf{x},\mathcal{M}_i)$$

Schwarz (1978) and Kass and Wasserman (1995) showed that if we use a certain "unit information" prior for the coefficients, then $\Pr(\mathbf{y}|\mathbf{x},\mathcal{M}_i)$ can be well approximated by $\exp(\frac{1}{2}BIC)$, where

$$BIC = -2\ell(\mathbf{y}|\mathbf{x},\boldsymbol{\beta}_{\mathcal{M}_i}) - \ln(N)d_{in} \tag{6}$$

Thus if we assume equal prior probabilities for all models, the model with the highest posterior probability is the one with lowest BIC. If we do not assume equal prior probabilities for all models, then $\exp(\frac{1}{2}BIC)$ approximates a "Bayes factor", i.e., a measure of the extent of support provided by the data itself for the given model, rather than a posterior probability; but this can still be used to guide model choice in the same way.

BIC is usually preferred over AIC (see, e.g., Rust, Simester, Brodie, and Nilikant, 1995) because it is a "consistent" model selection technique, unlike

AIC which tends to overfit and select too many variables. That is, assuming that there is a fixed finite number of models available and that one of them is the true model, then as the sample size gets large enough, the lowest-BIC model will be the true model, with probability approaching 100%. However, critics of BIC (see Shibata, 1989; Weakliem, 1999; Hastie et al., 2001; Burnham and Anderson, 2002; Leeb and Pötscher, 2005) suggest that its asymptotic properties are based on unrealistic assumptions and that BIC may prefer overly simple models for modest sample sizes. This debate is reviewed in Shao (1997), Kuha (2004) and Yang (2004, 2005). In simulations, BIC has usually performed better than AIC, unless the true model has many very small coefficients (see Dziak, Li, and Collins, 2005). Practical applications of BIC are explored in depth by Raftery (1995), Hauser (1995) and Wasserman (2000).

2.2 Continuous Penalties

The methods described so far are discontinuous. For some κ_j, if $\hat{\beta}_j = \kappa_j + .001$ the coefficient will not shrink, but if $\hat{\beta} = \kappa_j - .001$ it will be set to zero. Thus the final selected estimate $\hat{\beta}$ is not a continuous function of the full-model least-squares estimate $\hat{\beta}$, and not a continuous function of the original data. This is problematic, because a small change in the original data might cause a different model to be selected. This uncertainty may lead to bias, as well as to instability, i.e., unexpectedly high sampling variance relative to the model size selected (see Breiman, 1996; Zucchini, 2000; Miller, 2002). The problem of instability is especially severe in the presence of multicollinearity, since here the estimate for one coefficient strongly affects the estimates for the others. Some of the error variance reduction which we would expect to gain by using a smaller submodel instead of the full model is lost again due to the fallible nature of the data-driven choice of submodel (Thompson, 1995; Breiman, 1996; Babyak, 2004).

Besides potentially causing poorer estimation performance, the discontinuity of the penalty is also responsible for the computational difficulties associated with variable selection when p is high. The $\hat{\beta}$ for, say, the best-AIC model cannot be found from the full model by a Newton-Raphson or similar algorithm; it must be found by fitting each candidate subset separately and checking which one has the best AIC. Unless d is very large (say > 50), this is not a problem for ordinary linear regression, because of the availability of fast computers and specialized algorithms such as that of Furnival and Wilson

(1974). However, it remains problematic for more complicated models or high d.

One approach to making selection more continuous, in the hopes of reducing sampling variability, is to adopt a more sophisticated complexity penalty than λd_{in}. In (1), we let $\mathcal{P}(\boldsymbol{\beta})$ be a continuous function rather than a discrete count of nonzero elements in $\boldsymbol{\beta}$.

2.2.1 Ridge Regression

Recall that in ridge regression (Hoerl and Kennard 1970; see Sen and Srivastava 1990 or Neter et al Neter, Kutner, Nachtsheim, and Wasserman 1996) we minimize RSS subject to the constraint $\sum_{j=1}^{d} \boldsymbol{\beta}_j^2 < \tau$, by setting

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda_n \mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

where $\lambda_n \geq 0$ is a constant. Ridge regression does not accomplish subset selection; although it reduces the absolute values of the individual coefficients, it does not set any to zero. However, it can improve the predictive ability of a model by reducing sampling variance. It reduces the tendency to overfit when there are many predictors or when the predictors are highly collinear. Both ridge and subset selection are forms of "regularization" : they constrain the model somehow in an attempt to stabilize estimation and achieve a favorable bias-variance tradeoff (see Breiman, 1996; Fu, 1998). Intuitively, both subset selection and ridge regression favor a fit with a "small" $\boldsymbol{\beta}$; but subset selection favors a small *number* of nonzero coefficients, while ridge regression favors small *values* for the coefficients themselves. Specifically, ridge regression constrains the squared L_2 norm $\sum_{j=1}^{d} \left|\boldsymbol{\beta}_j\right|^2$, while subset selection constrains the L_0 norm $\sum_{j=1}^{d} \left|\boldsymbol{\beta}_j\right|^{0+}$, where we define f^{0+} as $I\{f \neq 0\}$ (see Frank and Friedman, 1993; Fu, 1998) and $I(\cdot)$ is the indicator function.

2.2.2 LASSO

Tibshirani (1996) proposed that a penalty based on the L_1 norm $\sum_{j=1}^{d} \left|\boldsymbol{\beta}_j\right|$ can to some extent combine the benefits of the L_0 norm (automatic deletion of small coefficients) with those of the L_2 norm (stabilization by shrinkage). His Least Absolute Shrinkage and Selection Operator (LASSO) is simply penalized least-squares with the L_1 penalty:

$$\text{RSS} + \lambda \sum_{j=1}^{d} \boldsymbol{\beta}_j^2 \tag{8}$$

Tibshirani proposed the LASSO as an improvement upon an earlier, more complicated constrained least squares proposal, the "garotte" of Breiman (1995). The main advantage of the LASSO over ridge regression is that LASSO automatically shrinks small coefficients to zero, thus simplifying the model (see Tibshirani, 1996; Fu, 1998; Li, Dziak, and Ma, 2006, for why this occurs for L_1 and not L_2). For ease of computation, LASSO is in between ridge (easy) and the L_0 penalties (difficult due to the need to consider many separate models). Although the LASSO criterion cannot be minimized explicitly, it can be minimized by linear programming (Tibshirani, 1996), by a modified Newton-Raphson method (see Fu, 1998; Ojelund, Madsen, and Thyregod, 2001; Fan and Li, 2001, and Section 2.2.4), or by the very efficient "least-angle" or LARS algorithm described by Efron, Hastie, Johnstone, and Tibshirani (2004) (see also Tibshirani, 2002).

Both ridge and LASSO regression can be viewed as Bayesian procedures, with priors on individual coefficients. Specifically, the ridge and LASSO solutions are modes of the posterior likelihood function, under normal or double-exponential priors, respectively, on the β_j's; the size of λ is inversely related to the assumed variance for the prior. Bayesian interpretations of LASSO are explored in Tibshirani (1996) and Yuan and Lin (2005).

Besides L_1, a continuum of penalties between L_0 and L_2 is possible. Frank and Friedman (1993), Tibshirani (1996), Fu (1998), and Fan and Li (2001) describe the properties of "bridge regression," with a penalty based on $\sum_{j=1}^{d} |\beta_j|^q$ with any q between 0 and 2. q could be chosen arbitrarily *a priori*, or could perhaps be optimized by some data-driven procedure. The behavior of the resulting estimate would be intermediate between subset selection (L_0) and ridge (L_2), i.e., with small q tending to give a few large coefficients and large q tending to give many small coefficients. However, it is difficult to choose q and there is no compelling advantage over simply using LASSO. The L_q penalties are reviewed in Fu (1998), Knight and Fu (2000) and Li et al. (2006).

L_1-penalized regression has had a very favorable reception within the statistical literature, because it provides a more computationally practical and mathematically elegant alternative to all-possible-regressions. However, some problems remain. Leng, Lin, and Wahba (2004) showed that the LASSO is not asymptotically consistent, in that no matter how large the sample size is,

there is still a nonnegligible chance of its choosing the wrong model (however, see Zhao and Yu, 2005, for further discussion). Results in Fan and Li (2001) and Li et al. (2006) suggest an intuitive explanation: roughly, since LASSO controls shrinkage and selection with a single tuning parameter λ, the user may be forced to choose between having overly shrunken estimates (high λ) and too many predictor variables (low λ). This problem is addressed by a newer penalty function, the SCAD.

2.2.3 SCAD

The LASSO provides model selection and a continuous estimator. However, it may shrink coefficients more than is desirable (i.e., introduce too much bias). This motivated Fan (1997) to propose the Smoothly Clipped Absolute Deviation (SCAD) penalty. SCAD penalizes coefficients differently based on their original size. This is a compromise between LASSO and ridge (which penalize all coefficients uniformly), and the L_0 selection methods (which delete small coefficients and keep large ones unshrunken, but sacrifice continuity and stability). These distinctions are illustrated in Table 1 and Figure 2 for the column-orthogonal case. Specifically, the SCAD penalty is defined as $\mathcal{P}(\boldsymbol{\beta}) = \sum_j p(|\beta_j|)$ such that

$$p_j(|\beta_j|) = \begin{cases} \lambda|\beta| & \text{if } 0 \leq |\beta| < \lambda \\ \frac{(a^2-1)\lambda^2-(|\beta|-a\lambda)^2}{2(a-1)} & \text{if } \lambda \leq |\beta| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| \geq a\lambda \end{cases} \tag{9}$$

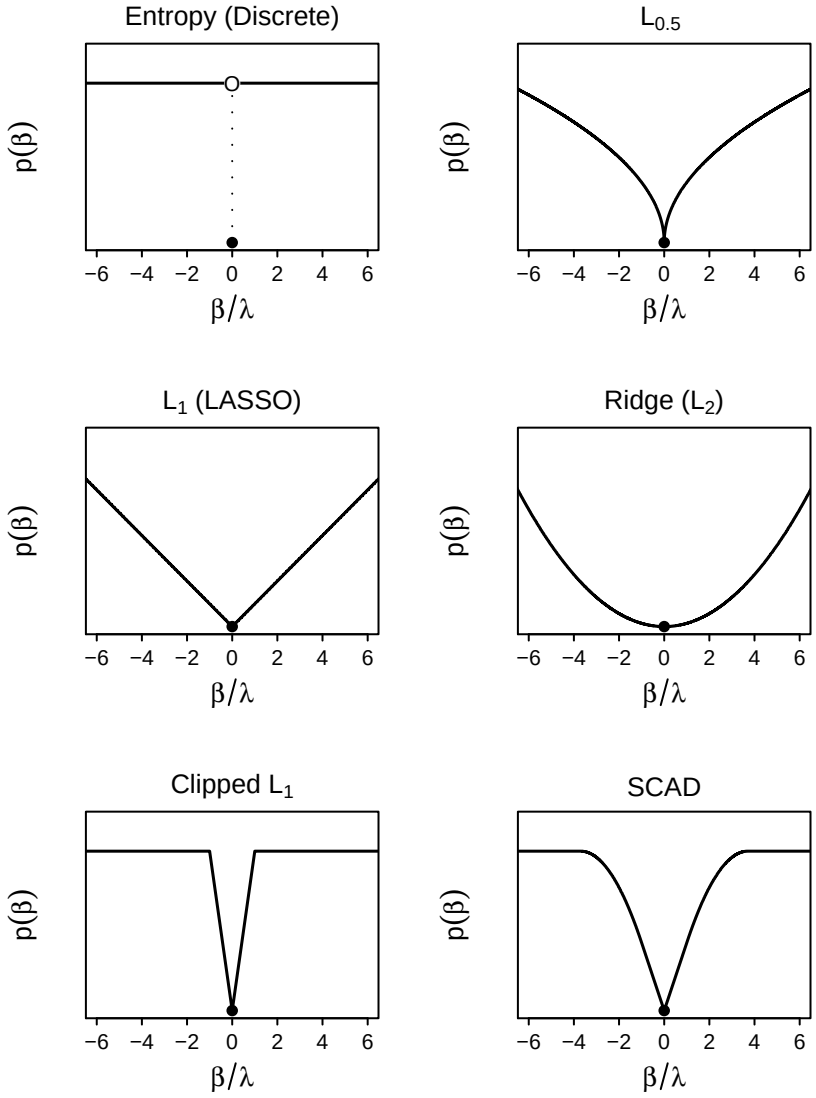
so

$$p'(|\beta_j|) = \begin{cases} \lambda & \text{if } |\beta_j| < \lambda \\ (a-1)^{-1}(a\lambda - |\beta_j|)) & \text{if } \lambda < |\beta_j| < a\lambda \\ 0 & \text{if } |\beta_j| > a\lambda \end{cases} \tag{10}$$

Compare LASSO, for which $p(|\beta_j|) = \lambda|\beta_j|$ and $p'(|\beta_j|) = \lambda$. The SCAD penalty function is shown in Figure 1, with the ridge and LASSO penalties also shown for comparison.

As $a \to \infty$, SCAD approaches the behavior of the LASSO. As $a \to 0$ the SCAD penalty would converge pointwise to the entropy penalty; however, a must be greater than 2 to provide a continuous estimator. a can be chosen using cross-validation or generalized cross-validation, or it can be set in advance to 3.7 (for $a = 3.7$ as approximately optimal see Fan and Li, 2001, p. 1350). The advantage of SCAD over LASSO is that SCAD can treat small

Figure 1: Ridge, Lasso, and SCAD Penalty Functions in the Orthogonal Case with λ=0.1



Notes. This figure shows the penalty functions discussed in this section, including the L_0 or entropy penalty (2), three cases of bridge regression (specifically $L_{0.5}$, shown only for comparison; L_1, which is LASSO for linear models; and L_2, which is ridge regression for linear models), and the smoothly clipped absolute deviation penalty (SCAD). The y axes are scaled to make the shapes of the functions easy to compare. Notice that the SCAD penalty starts out linear, then slows down and finally levels off and becomes flat; its boundedness suggests that past a certain point we do not care how big coefficients get. Also shown for comparison is a less sophisticated version of SCAD, mentioned in Antoniadis and Fan (2001): the (non-smoothly) clipped L_1 penalty $p_j(|\beta_j|) = \lambda \min(\lambda, |\beta_j|)$. The latter penalty could be made more like SCAD by generalizing it to $p_j(|\beta_j|) = \lambda \min(a\lambda, |\beta_j|)$, but it does not have a continuous first derivative as SCAD does.

and large coefficients separately. With SCAD, we penalize small coefficients proportionally heavily, to encourage their deletion. We penalize large coefficients proportionally lightly; since we are sure we do not want to delete them, we do not want to introduce unnecessary bias by shrinking them. We shrink intermediate coefficients somewhat, for the sake of stabilization and continuity.

Like the LASSO, the SCAD combines variable selection and regularization. However, SCAD can also provide a smaller bias in coefficient estimation than LASSO because it is bounded as a function of β. Furthermore, under certain conditions, including an appropriate choice of λ, the SCAD-penalized least squares estimator is a consistent (BIC-like) model selector, as described in Fan and Li (2001). With an appropriate choice of tuning parameter, SCAD has the "oracle property": assuming the true model is a subset of the full model available, it is consistent in the BIC sense, and so its asymptotic bias and variance are no higher than if the nuisance coefficients had not been in the model at all. The LASSO does not have this consistency (Fan and Li (2001), Leng et al. (2004)); for large λ it is too biased, and for small λ it overfits like AIC.

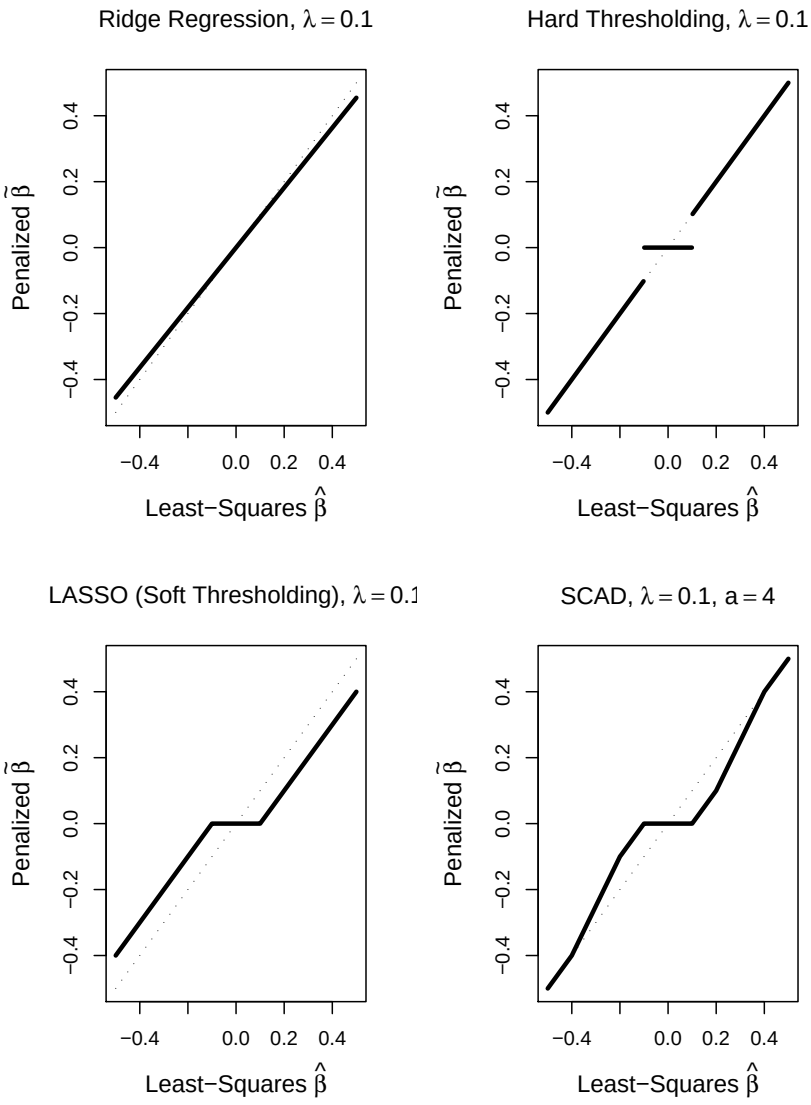
In the case of linear modeling with column-orthogonal design matrix (i.e., in which all predictors are uncorrelated and have mean zero and variance one, so $\mathbf{y}^T\mathbf{y} = n\mathbf{I}$) penalized criteria often have explicit solutions which make their properties more evident. These are described in Antoniadis and Fan (2001), Fan and Li (2001), and Li et al. (2006). Choose a tuning constant $\lambda_0 \geq 0$ and let $\lambda_n = N\lambda_0$. The multiplier N is needed so that the penalty term c can keep the same proportional size relative to the goodness-of-fit term ℓ in (1) as N grows. Then the penalized estimates $\hat{\beta}_j$ are simple functions of the least-squares estimates $\hat{\beta}_j$, as shown in Table 1 and Figure 2; see Hastie et al. (2001, p. 71), Fan and Li (2001), or Li et al. (2006) for details.

2.2.4 Local Quadratic Approximation (LQA) Algorithm

A general method of computing penalized estimates is obtained from a slight modification of the Newton-Raphson algorithm, using a local quadratic approximation (LQA; see Fan and Li, 2001; Hunter and Li, 2005; Li et al., 2006) for the penalty. This subsection describes LQA for the penalized likelihood case, but it is very broadly applicable and is used throughout this thesis with appropriate modifications to find penalized GEE and QIF estimates.

The LQA algorithm is motivated as follows. Let $\mathrm{L}(\beta)$ be a loss function

Figure 2: Some Penalized Estimators as Functions of the Least-Squares Estimator in the Orthogonal Case



Note. Each of these estimators, except SCAD, has some undesirable property. The ridge estimator, like the least-squares estimator itself, does not sets coefficients to zero so it does not give a sparse model. The hard-thresholding estimator (resulting from the L_0 penalty) is discontinuous. The LASSO introduces considerable bias. SCAD takes a more sophisticated approach: sets small coefficients to zero (for sparsity), shrinks medium-sized coefficients (for stability and continuity), and leaves large coefficients relatively unchanged (to reduce modeling bias).

Table 1: Orthogonal Case Behavior of Other Selection and Regularization Methods

	Estimate in Orthogonal-Predictors Linear Model Case
Classical	$\tilde{\beta}_j = \begin{cases} 0 & \text{if } \left\|\hat{\beta}_j\right\| < \lambda_0 \\ \hat{\beta}_j & \text{if } \left\|\hat{\beta}_j\right\| > \lambda_0 \end{cases}$
Ridge	$\tilde{\beta}_j = (1+\lambda_0)^{-1}\hat{\beta}_j$
LASSO	$\tilde{\beta}_j = \begin{cases} \hat{\beta}_j + \lambda_0 & \text{if } \hat{\beta}_j < -\lambda_0 \\ 0 & \text{if } \left\|\hat{\beta}_j\right\| < \lambda_0 \\ \hat{\beta}_j - \lambda_0 & \text{if } \hat{\beta}_j > \lambda_0 \end{cases}$
SCAD	$\tilde{\beta}_j = \begin{cases} 0 & \text{if } \left\|\hat{\beta}_j\right\| \le \lambda_0 \\ \hat{\beta}_j - \lambda_0 \operatorname{sgn}\left(\hat{\beta}_j\right) & \text{if } \lambda_0 < \left\|\hat{\beta}_j\right\| < 2\lambda_0 \\ \left\{\frac{a-1}{a-2}\hat{\beta}_j - \frac{a\operatorname{sgn}(\hat{\beta}_j)}{a-2}\lambda_0\right\} & \text{if } 2\lambda_0 < \left\|\hat{\beta}_j\right\| < a\lambda_0 \\ \hat{\beta}_j & \text{if } \left\|\hat{\beta}_j\right\| > a\lambda_0 \end{cases}$

Note. The "classical" criterion is here parameterized as $\ell(\boldsymbol{\beta}) - n\frac{1}{2}\lambda_0^2 d_{in}$ so that λ is the threshold parameter; see Foster and George (1994), Fan and Li (2001), and Li et al. (2006). λ here is $\sqrt{2/\sigma}\lambda$ in the notation of (2), so that, say, AIC corresponds to $\lambda = 2/\sqrt{\sigma}$.

of interest, e.g., a negative log-likelihood or log-quasi-likelihood or a sum of squared errors; assume it is a nonnegative and twice differentiable function of β whose expected value is uniquely minimized at the true value β_0 . We could estimate β_0 by minimizing $\mathrm{L}(\beta)$. However, in more generality suppose we want to minimize a penalized form $\mathrm{PL}(\beta)$, where

$$\mathrm{PL}(\beta) = \mathrm{L}(\beta) + N\sum\nolimits_{j=1}^{d} p_\lambda(|\beta_j|) \tag{11}$$

Let β_{opt} be the unknown true minimizer of $\mathrm{PL}(\beta)$, i.e., the solution to $E(\partial\,\mathrm{PL}\,/\partial\beta) = \mathbf{0}$. Choose some starting value $\beta^{(0)}$ near β_{opt}. Fan and Li (2001) suggest approximating $p_\lambda(|\beta_j|)$ by:

$$p_\lambda(|\beta_j|) \approx p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + \frac{1}{2}\left(p'_\lambda\left(\left|\hat{\beta}_j^{(0)}\right|\right) / \left|\hat{\beta}_j^{(0)}\right|\right)\left(\beta_j^2 - (\hat{\beta}_j^{(0)})^2\right) \tag{12}$$

Here $p_\lambda\left(\left|\beta_j^{(0)}\right|\right)$ can be treated as constant since it depends only on $\hat{\beta}^{(0)}$.

Then up to an additive constant we have

$$n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \approx \frac{1}{2}\ N\ \beta^T \mathbf{\Delta}^{(0)} \beta \tag{13}$$

where

$$\mathbf{\Delta}^{(0)} = \text{diag}\left(p'(\left|\hat{\beta}_1^{(0)}\right|)/\left|\hat{\beta}_1^{(0)}\right|, \ldots, p'(\left|\hat{\beta}_d^{(0)}\right|)/\left|\hat{\beta}_d^{(0)}\right|\right). \tag{14}$$

For L we can apply the usual Taylor approximation

$$\mathrm{L}(\beta) \approx \mathrm{L}(\beta^{(0)}) + \dot{\mathrm{L}}(\beta^{(0)})^T(\beta - \beta^{(0)}) + \frac{1}{2}(\beta - \beta^{(0)})^T \ddot{\mathrm{L}}(\beta^{(0)})(\beta - \beta^{(0)}) \tag{15}$$

so that up to a constant

$$\begin{aligned} \mathrm{PL}(\beta) \approx\ & \mathrm{L}(\beta^{(0)}) + \dot{\mathrm{L}}(\beta^{(0)})^T(\beta - \beta^{(0)}) \\ & + \frac{1}{2}(\beta - \beta^{(0)})^T \ddot{\mathrm{L}}(\beta^{(0)})(\beta - \beta^{(0)}) + \frac{1}{2}\ N\ \beta^T \mathbf{\Delta}^{(0)} \beta \end{aligned} \tag{16}$$

where the right-hand side is a quadratic function of β. For convenience write $\mathrm{L}(\beta^{(0)})$ as $\mathrm{L}^{(0)}$. Then at $\dot{\mathrm{L}} = 0$,

$$\beta - \beta^{(0)} \approx -\left(\ddot{\mathrm{L}}^{(0)} + n\mathbf{\Delta}^{(0)}\right)^{-1}\left(\dot{\mathrm{L}}^{(0)} + N\mathbf{\Delta}^{(0)}\beta^{(0)}\right) \tag{17}$$

This suggests an iterative algorithm in which at every step of the algorithm we set $\beta^{(k)} = \beta^{(k-1)} - \delta^{(k)}$, where

$$\delta^{(k)} = \left(\ddot{\mathrm{L}}^{(k-1)} + N\ \mathbf{\Delta}^{(k-1)}\right)^{-1}\left(\dot{\mathrm{L}}^{(k-1)} + N\mathbf{\Delta}^{(k-1)}\beta^{(k-1)}\right) \tag{18}$$

Upon convergence, use the solution $\beta^{(k)}$ as as the estimator of β_{opt}. This very general algorithm was described by Fan and Li (2001).

One inconvenience in using the LQA algorithm is that $\mathbf{\Delta}^{(k-1)}$ is undefined if any of the $\beta_j^{(k-1)}$ is zero, and is numerically unstable if any of the $\beta_j^{(k-1)}$ is near zero. Fan and Li (2001) suggested discarding, at each step of the algorithm, all coefficients whose absolute value is less than some cutoff (e.g., .001). These coefficients are estimated as zero, and corresponding rows $\mathbf{X}$, $\beta^{(k-1)}$, $\beta^{(k)}$, and $\mathbf{\Delta}^{(k-1)}$ are removed to avoid numerical instability. Other approaches include perturbing the denominator in (12) by a small constant (Hunter and Li, 2005) or using an iterative conditional minimization algorithm (Li et al., 2006).

2.2.5 Estimating the Standard Error

Fan and Li (2001) recommend a sandwich estimator for the covariance matrix of the nonnull predictors. Let $\hat{\beta}^* = (\hat{\beta}_1^*, ..., \hat{\beta}_k^*)^T$ be the nonnull predictors. Then based on the Taylor expansion of Q in the LQA algorithm,

$$\text{Cov}(\hat{\beta}) \approx \left\{\ddot{\text{L}}(\hat{\beta}) + N\boldsymbol{\Delta}\right\}^{-1} \widehat{\text{Cov}}(\dot{\text{L}}(\hat{\beta})) \left\{\ddot{\text{L}}(\hat{\beta}) + N\boldsymbol{\Delta}\right\}^{-1} \tag{19}$$

where $\boldsymbol{\Delta} = \text{diag}\left(p'(\left|\hat{\beta}_1\right|)/\left|\hat{\beta}_1\right|, \ldots, p'(\left|\hat{\beta}_d\right|)/\left|\hat{\beta}_d\right|\right)$. $\widehat{\text{Cov}}(\dot{\text{L}}(\hat{\beta}))$ could be a model-based estimate (e.g., $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})$ in linear regression) or the more robust empirical estimate $\frac{1}{n}\sum \mathbf{q}_i\mathbf{q}_i^T$ where $\mathbf{q}_i$ is $\dot{Q}(\hat{\beta}^*)$ evaluated with $\mathbf{x}_i$ and $\mathbf{y}_i$. In the normal likelihood case with a model-based covariance estimate, (19) becomes

$$\widehat{\text{Cov}}(\hat{\beta}) = \hat{\sigma}^2 \left\{\mathbf{X}^T\mathbf{X} + N\boldsymbol{\Delta}\right\}^{-1} (\mathbf{X}^T\mathbf{X}) \left\{\mathbf{X}^T\mathbf{X} + N\boldsymbol{\Delta}\right\}^{-1} \tag{20}$$

2.2.6 Selecting the Tuning Parameter

Penalized likelihood methods require selection of the tuning parameter λ. Some rule is needed so that λ can be selected in a principled way. Generally, some data-driven method such as cross-validation or generalized cross-validation is used (see Fan and Li, 2001).

For normal linear models, Generalized Cross-Validation (GCV) is a C_p-like model fit statistic used to conveniently approximate the predictive sum of squared errors which would be obtained from leave-one-out cross-validation (see Craven and Wahba, 1979; Hastie et al., 2001). For linear smoothers (i.e., for which $\hat{\mathbf{y}} = \mathbf{M}\mathbf{y}$), GCV is

$$GCV\,(\lambda) = \frac{1}{N}\frac{\text{RSS}(\boldsymbol{\beta}(\lambda))}{(1 - N^{-1}\text{df}(\lambda))^2} \tag{21}$$

where the effective number of parameters $\text{df}(\lambda)$ is the trace of the smoothing (i.e., projection or "hat") matrix $\mathbf{M}$; e.g., for ridge regression $\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$ since $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. LASSO and SCAD are not linear smoothers, but in the Gaussian case they can be viewed as approximately linear. At a given step of the LQA algorithm we set $\beta^{(m)} = (\mathbf{X}^T\mathbf{X}+N\boldsymbol{\Delta})^{-1}\mathbf{X}^T\mathbf{y}$, so Tibshirani (1996) and Fan and Li (2001) suggested using $df_T = \text{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X}+n\boldsymbol{\Delta})^{-1}\mathbf{X}^T)$ as a degrees of freedom estimate. This is discussed further in Li et al. (2006).

For linear models, GCV is asymptotically equivalent to C_p, AIC, and

leave-one-out cross-validation (see Shao, 1997; Hastie et al., 2001). Thus, Wang, Li, and Tsai (2005) argued that since GCV will tend to overfit, the performance of SCAD may be improved by using a modified criterion similar to BIC. Minimizing $GCV\,(\lambda)$ is equivalent to minimizing the AIC-like criterion $\log(n^{-1}RSS(\boldsymbol{\beta}(\lambda))) + \mathrm{df}(\lambda)$. Wang et al. (2005) suggest instead using

$$BIC(\lambda) = \log(\frac{\mathrm{RSS}(\boldsymbol{\beta}(\lambda))}{N}) + \frac{\log(N)}{N}\mathrm{df}(\lambda)$$

a variation of (6) taking shrinkage into account. A similar proposal was also made by Zou, Hastie, and Tibshirani (2004) for the LASSO. The GCV and BIC criteria can be extended beyond linear models; we could replace RSS with a weighted sum of squares, or a Poisson or binomial deviance, and adjust the formulas accordingly.

The degrees of freedom of the model can be defined in several ways besides df_T. The easiest approach is to ignore shrinkage and use $df_s \equiv d_{in}$; Zou et al. (2004) show that this is reasonable despite its simplicity. Fu (2003) took both shrinkage and deletion into account, in a somewhat ad hoc way, and let

$$df_F \equiv d_{in}\frac{\left\|\hat{\boldsymbol{\beta}}_{\mathrm{shrunken}}\right\|_{L_1}}{\left\|\hat{\boldsymbol{\beta}}_{\mathrm{unshrunken}}\right\|_{L_1}}$$

Another approach is to measure the effect of the penalty on the Hessian matrix or perhaps the covariance matrix, using, e.g.,

$$df_H \equiv \ddot{\ell}(\boldsymbol{\beta})\left(\ddot{\ell}(\boldsymbol{\beta}) + N\boldsymbol{\Delta}\right)^{-1} \tag{22}$$

For linear models some of these formulas are equivalent; $df_H = df_T$ since

$$\mathrm{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X}+N\boldsymbol{\Delta})^{-1}\mathbf{X}^T\right) = \mathrm{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X}+N\boldsymbol{\Delta})^{-1}\mathbf{X}^T\right), \tag{23}$$

and of course, without shrinkage (i.e., if $\boldsymbol{\Delta} = \mathbf{0}$), $df_H = df_T = df_s$. We would expect that df_H, df_T, and df_F would usually behave similarly to each other, and that df_s would not be much different but would lead to a somewhat heavier penalty (Li et al., 2006).

3 Variable Selection for Penalized GEE

This section reviews and compares proposals for extending penalized-likelihood-type model selection criteria for use with GEE modeling. Let us start by reviewing GEE itself and establishing notation.

Suppose that for each of n subjects $i = 1, ..., n$, we take J_i univariate observations $y_{i1}, ..., y_{iJ_i}$ at times t_{ij}. Along with each observation y_{ij}, there are observed d covariates $x_{ij1}, ..., x_{ijd}$. Let $N = \sum_i J_i$ be the number of observations; note that in the previous section $n = N$ but in this and the following sections $n < N$.

Suppose that our interest is in modeling the marginal population-level relationships of these covariates with the response, rather than predicting the responses of given subjects, so that we are not precisely modeling within-subject correlation. We would like to fit a linear (or generalized linear, according to the form of the response variable; see McCullagh and Nelder, 1991) model, but this usually requires the assumption of independent observations, unlike the current situation in which each subject is a cluster of intercorrelated observations. Ignoring the correlation among different observations from the same individual could lead to inefficient estimation of the regression coefficients and underestimation of standard errors. The generalized estimating equations (GEE) approach, proposed by Liang and Zeger (1986), extends generalized linear modeling (GLIM) to longitudinally clustered data.

In GLIM, it is assumed that univariate observations Y_i are taken independently from an exponential-family distribution depending on known covariates $\mathbf{x}_i$ and unknown coefficients β through a known link function $\mathrm{g}(\cdot)$, such that the mean is $\mu_i = \mathrm{g}(\mathbf{x}_i\beta)$, and the variance is $\mathrm{Var}(Y_i) = \phi\,\mathrm{V}(\mu_i)$. For example, in ordinary linear modeling $\mathrm{g}(\nu) = \nu$, $\phi = \sigma^2$, and $\mathrm{V}(\mu_i) = 1$; in logistic regression $\mathrm{g}(\nu) = \exp(\nu)/(1 + \exp(\nu))$, $\phi = 1$, and $\mathrm{V}(\mu_i) = (\mu_i)(1 - \mu_i)$. The exponential family assumption allows GLIM estimates to be considered maximum-likelihood estimates; without it, the GLIM estimate is a maximum quasi-likelihood estimate (see McCullagh and Nelder, 1991; Agresti, 2002). The GLIM estimate is obtained by solving the (quasi-)score equation

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = 0 \tag{24}$$

for β, where $\mathbf{y} = [y_1, ..., y_n]^T$ is the vector of observed data, $\mu = [\mu_1, ..., \mu_n]^T$ is the vector of means $\mu_i = \mathrm{g}^{-1}(\mathbf{x_i}\beta)$, $\mathbf{D} = \partial\,\mathrm{g}^{-1}(\mathbf{x}_i\beta)/\partial\beta$ is a matrix such that $\dot{\mu}_{ik} = \partial\boldsymbol{\mu}_i(\mathbf{x}_i\beta)/\partial\beta_k$, and $\mathbf{V} = \mathrm{Cov}(\mathbf{Y})$. In practice, (24) is solved by alter-

nately estimating $\mathbf{V}$ and setting the resulting quantity to zero; for the normal distribution this would be the common method of iteratively reweighted least squares (see Dunlop, 1994).

Under the usual assumption of independence of observations, $\mathbf{V}$ is a diagonal matrix with $\mathbf{V}_{ii} = \phi \, \mathrm{V}(\mu_i)$ (or, absorbing the scale parameter into V, just $\mathbf{V}_{ii} = \mathrm{V}(\mu_i)$). GEE extends GLIM by allowing $\mathbf{V}$ to be block-diagonal rather than diagonal, i.e., subjects are independent but observations within the same subject or cluster are not independent (see Liang and Zeger, 1986; Dunlop, 1994). We now decompose $\mathbf{V}$ into marginal variance and correlation structure, and also rewrite the left-hand side of (24) as a sum of blocks. Then (24) becomes the "generalized estimating equations"

$$\sum\nolimits_{i=1}^{n} \mathbf{D}_i^T \mathbf{A}_i^{-1/2} \mathbf{R}_i^{-1} \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \tag{25}$$

where $\mathbf{x}_{ij}$ is the d-vector of covariates for the i^{th} subject at the j^{th} time; $\mathbf{y}_i = [y_{i1}, ..., y_{iJ_i}]^T$ is the J_i-vector of observed data for the i^{th} subject; $\boldsymbol{\mu}_i = [\mathrm{g}^{-1}(\mathbf{x}_{i1}\beta), ..., \mathrm{g}^{-1}(\mathbf{x}_{in}\beta)]^T$; $\mathbf{D}_i$ is a matrix such that $\mathbf{D}_{ijk} = \partial \boldsymbol{\mu}_i(\mathbf{x}_{ij}\beta)/\partial \beta_k$; $\mathbf{A}_i$ is a diagonal matrix such that $(\mathbf{A}_i)_{jj} = \phi \, \mathrm{V}(\boldsymbol{\mu}_i)$; and $\mathbf{R}_i$ is the working within-subject correlation matrix. If the $\mathbf{R}_i$ are correctly specified, then $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2} = \sigma_i \equiv \mathrm{Cov}(\mathbf{Y}_i)$. For models using the canonical link (see McCullagh and Nelder, 1991), $\mathbf{D}_i = \mathbf{A}_i \mathbf{X}_i$. For linear models, $\mathbf{A} = \mathbf{I}$.

Liang and Zeger (1986) suggested approximating $\mathbf{R}$ by a working correlation matrix $\hat{\mathbf{R}}$ involving only one or a few nuisance parameters α, using *ad hoc*, method-of-moments-like estimators for these α. They showed that an incorrect choice of working correlation structure leads only to suboptimal efficiency, while still allowing consistent and asymptotically normal estimation of β. Many working covariance structures are possible, but for longitudinal data a few simple choices are common: often a first-order autoregressive (AR(1); i.e., $\mathrm{Cov}(y_i, y_j) \propto \alpha^{|t_i - t_j|}$) correlation for time series, or exchangeable correlation (i.e., equicorrelated or compound symmetric; $\mathrm{Cov}(y_i, y_j) = \alpha \; \forall i \neq j$) for other cases of clustered or nested data (see Liang and Zeger, 1986; Diggle, Heagerty, Liang, and Zeger, 1998). Each of these modeling choices involves only one or a few nuisance parameters and suggests obvious moment estimators for them (see Liang and Zeger 1986, pp. 17-18, or SAS Institute, Inc. 2004, p. 1674).

3.1 Model Selection for Generalized Estimating Equations

Until recently, the only model selection approach established for GEE was sequential testing, typically with Wald z-tests on individual coefficients. However, testing alone is sometimes too simplistic for choosing predictive models (see Thompson, 1995; Burnham and Anderson, 2002; Cantoni, Flemming, and Ronchetti, 2005; Dziak et al., 2005; Gurka, 2006). Penalized model fit statistics like AIC are useful because they can be used to compare nonnested models, and can also be used to find lists of plausible models, whereas sequential testing provides only a single supposedly best answer (see Neter et al., 1996; Ramsey and Schafer, 2002; Cantoni et al., 2005; Dziak et al., 2005).

One simple measure of model fit for GEE models is marginal R^2 (Zheng, 2000; Ballinger, 2004), an extension of classical R^2 defined as

$$R^2_m = 1 - \frac{\sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2}{\sum_i \sum_j (y_{ij} - \bar{y})^2} \tag{26}$$

where $\hat{y}_{ij}$ is the marginal expected value from the model, given covariates $\mathbf{x}_{ij}$ (e.g., $\mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}$ for the linear model), and $\bar{y}$ is the grand mean of all obvervations.

This measure ignores correlation and does not attempt to weight the residuals, even though $\hat{\boldsymbol{\beta}}$ comes from a model with working covariance weights; this may not be optimal but is very helpful as a simplification. A more serious problem is that, although R^2_m could be useful for certain selection tasks (such as choosing a working correlation structure), it cannot generally be used for variable selection purposes, since like classical R^2 it would always lead to choosing the largest model available.

Fortunately, there has recently been significant work on extending other classic model selection criteria – including C_p, AIC, BIC, LASSO, and SCAD – to cluster-correlated data (see Dziak and Li, 2006), as outlined below.

3.1.1 Generalizing Mallows' C_p

Cantoni et al. (2005) suggested a generalization of Mallows' C_p (Mallows, 1973) for GEE models, for estimating predictive risk under a general weighted loss function. The weights can be adjusted in order to account for correlation within subjects as well as downweighting unusual observations, potentially providing robustness against outliers and model misspecification. They derive a GC_p statistic to estimate the resulting risk function. The resulting

statistic is complicated and may require Monte Carlo approximation to evaluate. However, in the case of no special weighting for robustness, the GC_p has the form

$$GC_p = \sum_{i=1}^{n}\sum_{j=1}^{n_i} r_{ij}^2 - \sum_{i=1}^{n} J_i + 2\,df_C. \tag{27}$$

where $r_{ij} = (\mathbf{A}_i)_{jj}^{-1}(y_{ij} - \mu_{ij})$, $df_C = \text{tr}(\mathbf{H}^{-1}\mathbf{Q})$, $\mathbf{H} = n^{-1}\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i$, $\hat{y}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$, and
$\mathbf{Q} = n^{-1}\sum_{i=1}^{n}\mathbf{D}_i^T\mathbf{A}_i^{-1}\mathbf{D}_i$ (see Dziak and Li, 2006).

3.1.2 Generalizing the AIC

Pan (2001) considered the problem of extending the classical derivation of AIC (Akaike, 1973), which involves estimating the relative Kullback-Leibler discrepancy of each likelihood model from an unknown true model, to a GEE setting. In GEE, the likelihood is not specified, but a quasi-likelihood (see Wedderburn, 1974; McCullagh and Nelder, 1991; Agresti, 2002) may be implicitly specified. Pan adapted the original derivation of the AIC, which estimated of an expected model log-likelihood under the true model, to instead estimate an expected working-independence quasi-likelihood. This resulted in the quasi-AIC

$$QIC = -2QL(\hat{\beta}_{Ind}) + 2\,\text{tr}(df_P)) \tag{28}$$

where $df_P = \mathcal{I}_{Ind}\widehat{\text{Cov}}(\hat{\beta})$, $\mathcal{I}_{Ind}$ is $-\,E\frac{\partial^2}{\partial\beta\beta^t}\ell_{Ind}(\beta)\Big|_{\beta=\hat{\beta}_{Ind}}$, i.e., the Fisher information under independence, and $\widehat{\text{Cov}}(\hat{\beta})$ is the robust (GEE) sandwich estimate. For the quasi-likelihood QL we can often use the likelihood ℓ_{Ind} under working independence; if overdispersion exists then we incorporate the scale parameter and use $\ell_{Ind}(\beta)/\phi$. Note that in Dziak and Li (2006), functions like QL here were called "pseudo-quasi-likelihoods" because they combine the idea of a pseudo-likelihood (a function which could be a likelihood but is believed to be only an approximation rather than an accurate description of the actual probability structure which generated the data) and a quasi-likelihood (a function which generates a good estimating equation given an assumed mean and variance structure).

This QIC is similar to Takeuchi's information criterion, a more general form of AIC in the classical case (see Shibata, 1989; Burnham and Anderson, 2002). If the model is adequate and if the responses are all independent, then $\mathcal{I}_{Ind}\widehat{\text{Cov}}(\hat{\beta}) \approx df_s$ so $QIC \approx AIC$. However, the dependence of the QIC on working independence may impede its performance in cases of strong

correlation. The mathematical difficulties of trying to define QIC without working independence are described by Pan (2001). He suggests possibly using a naïve extension of AIC,

$$-2QL(\hat{\beta}_{(working)}) + 2df_s \tag{29}$$

for other working structures, but does not explore this.

3.1.3 Generalizing the BIC

In the spirit of (29), we could also propose various extensions for the BIC statistic, by penalizing some quasi-likelihood or pseudo-likelihood instead of a likelihood. These would probably remain consistent selectors like the original BIC under appropriate conditions, although they might not retain the Bayesian interpretation of the original. Let us consider some possibilities for doing so; for further discussion see Dziak and Li (2006) and Gurka (2006).

Suppose first that we wish to extend an AIC-like fitting criterion. It is convenient to consider AIC first because the AIC penalty constant is not a function of n. In the homoskedastic Gaussian case, if the within-cluster covariance matrices $\mathbf{\Sigma}_i$ are known, then the log-likelihood function is of course

$$\ell(\beta|\mathbf{X},\mathbf{y}) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|\mathbf{\Sigma}_i| - \frac{1}{2}\sigma^{-2}\text{WSS} \tag{30}$$

where $\text{WSS} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^T\mathbf{R}_{\tau i}^{-1}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})$ and $\mathbf{R}_{\tau i} = \text{Corr}(\mathbf{y}_i|\mathbf{x}_i)$. The difficulty with the GEE scenario is that the true $\mathbf{\Sigma}_i$, or more specifically the true $\mathbf{R}_{\tau i}$, are unknown; we only have the working structures $\mathbf{R}_i$. (Note that $\mathbf{A}_i^{1/2}\mathbf{R}_{\tau i}\mathbf{A}_i^{1/2} = \mathbf{\Sigma}_i$ but $\mathbf{A}_i^{1/2}\mathbf{R}_{\tau i}\mathbf{A}_i^{1/2}$ generally does not.)

This difficulty reminds us of a simpler problem in the classic (independent-errors) case, in which $\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$ but σ^2 is unknown. If we use the estimate from the model being fitted, i.e., $\sigma^2 = \text{RSS}/n$, we get the usual formula used with AIC in regression:

$$AIC = \text{constant} + N\log(\text{RSS}/N) + 2d_{in} \tag{31}$$

where $\text{RSS}=\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^T(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})$. If we instead use the estimate of σ^2 from the largest model available, we get

$$\text{constant} + \frac{1}{\hat{\sigma}^2}\text{RSS} \tag{32}$$

algebraically equivalent to Mallows' C_p. Both versions behave similarly asymptotically despite handling the nuisance parameter differently.

Returning to (30), the nuisance parameter is now $\mathbf{\Sigma} \equiv \mathrm{Cov}(\mathbf{y_i}|\mathbf{x_i})$ instead of σ^2. The simplest solution is to ignore correlation and simply use (31) or (32). That is, regardless of the working correlation we use for estimating the coefficients within each candidate model, we will always use working independence for evaluating model fit and choosing a model. This approach does not seem sensible, but as we will see, there is not much performance cost to using it.

Another approach is to use the same working correlation structure for fitting the model and for calculating AIC, i.e., to act entirely as in the classical situation, except that ℓ is now as a pseudo-likelihood rather than the true likelihood function. Thus we get AIC by plugging $\mathbf{V}_i$ into (30) in place of $\mathbf{\Sigma}_i$. One complication is that now $\mathbf{R}$ affects the fit criterion not only through WSS but also through the determinant term, which is different for each subset we consider, since each model gives a different estimate of the nuisance correlation parameters. Our new AIC is no longer as intuitive a measure of fit as (31) was, since the metric for measuring fit seems to change with each subset. One possibility for dealing with this is to drop the determinant term; this was done for convenience in the simulations in Dziak and Li (2006). We thus get a simple quadratic function of the residuals, but no longer exactly a penalized multivariate normal pseudo-likelihood. Another possibility, in the spirit of (32), would be to get an estimate of $\mathbf{\Sigma}_i$ (either model-based or robust) from the largest model available, and then treat this as the true $\mathbf{\Sigma}_i$ for evaluating all simpler models.

The following formulas are therefore available:

$$IC = N \log(\mathrm{RSS}/N) + c_n df_s \tag{33}$$

$$IC = N \log(\mathrm{WSS}/N) + c_n df_s \tag{34}$$

$$IC = N \log(\mathrm{WSS}/N) + c_n df_s - \sum_{i=1} n \log\det(\mathbf{R}_i) \tag{35}$$

$$IC = N \log(\mathrm{FWSS}/N) + c_n df_s \tag{36}$$

$$IC = \sigma^{-2}\mathrm{RSS} + c_n df_s \tag{37}$$

$$IC = \sigma^{-2}\mathrm{FWSS} + c_n df_s \tag{38}$$

where

$$\sigma^2 \text{ is the scalar variance estimate from the largest model,} \tag{39}$$

$$\text{WSS} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^T \mathbf{R}_{\text{working}_i}^{-1} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) \tag{40}$$

$$\text{FWSS} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^T \mathbf{R}_{\text{full model}_i}^{-1} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}), \text{ and} \tag{41}$$

$$c_n = 2 \text{ for AIC, or some } \log(\hat{n}) \text{ for BIC} \tag{42}$$

Thus we have at least $6 \times 4 = 24$ possibilities which need to be explored in some way. Also, we must choose whether to use the $\hat{\boldsymbol{\beta}}$ from the working-independence model, or from the chosen working model (e.g., AR(1)); in this thesis I use the latter, although Pan (2001) used the former. We would hope that either the exact choice of formula does not matter very much, or else that there is a clear winner, since otherwise the situation of an analyst wishing to interpret these criteria in practice could be very confusing. Fortunately, the former seems to be the case, as shown in Section 3.1.4.

Further complications arise if we wish to generalize BIC, since c_n now depends on sample size, which is ambiguously defined (unlike in the classical case, the number of subjects is no longer the same as the number of observations; see Raftery, 1995). In place of the classic penalty constant $\log(n) = \log(N)$ we must now use $\log(\hat{n})$ where $\hat{n}$ is some more general measure of sample size between n and N. Dziak and Li (2006) considered both a "light" BIC or BIC_1 with $\hat{n} = n$, and a "heavy" BIC or BIC_2 with $\hat{n} = N$. They found that although BIC_2 was more reasonable in light of (30), BIC_1 tended to give better performance than the latter in simulations, because BIC_2 often selected too small a model. Pauler (1998) distinguished between variables which are constant for each subject (e.g., gender, treatment group; for these the sample size is effectively n) and those which vary within subjects (e.g., time, temperature, mood; for these the effective sample size may be N). Similarly, Fu (2003) suggested an "effective sample size," defined in (49), which tries to account for the amount of within-subject correlation. These versions should all have the same asymptotic behavior if the J_i are uniformly bounded, since n and N are of the same order.

3.1.4 Simulations on the AIC and BIC Formulas

To investigate how much the choice of formula matters to performance, some simulation experiments were done. For each of two true correlation structures (AR(1) or compound symmetric), 100 datasets were generated, each consisting of $n = 50$ subjects with $J = 7$ observations each. Responses y_{ij} were generated as $y_{ij} = 20 + \mathbf{x}_{ij}\boldsymbol{\beta} + 6\epsilon_{ij}$, with $\boldsymbol{\epsilon}_i$ multivariate normal with either AR(1) or exchangeable true correlation, correlation parameter $\rho_y = .7$, and $\sigma_y^2 = 1$. The ten predictor variables x_{ijk} were multivariate normal with marginal mean 0, marginal variance 1, and exchangeable (compound symmetric) correlation with $\rho_x = .6$. The true regression parameters were $[3, 1, 0, 0, 2, 0, 0, 0, .5, 0]^T$. This scenario was designed to present a very difficult selection task with unfavorable signal-to-noise ratio (error variance of 36), in hopes that this extreme case would make differences among the formulas more clearly evident. In particular, the very small ninth coefficient will be deleted by overly parsimonious formulas, and some of the six null coefficients will be included by overly liberal formulas.

Results are shown in Tables 2 and 3. The performance of each criterion was measured by mean model error, correct deletions, and wrong deletions. **Correct deletions** are defined as the average number (per simulation) of truly zero coefficients correctly estimated as zero, and **wrong deletions** are the average number of truly nonzero coefficients erroneously set to zero (roughly, Type Two errors). Because the true β is $[3, 1, 0, 0, 2, 0, 0, 0, .5, 0]^T$, up to 6 correct deletions and 3 wrong deletions are possible. Wrong inclusions (Type One errors) are not shown because of course, if a method has an average of d correct deletions, it has an average of $6 - d$ wrong inclusions. Model error is measured as $\text{ME}(\hat{\beta}) = \|\hat{\beta}\mathbf{X} - \beta\mathbf{X}\|^2 = (\hat{\beta} - \beta)^T E(\mathbf{x}\mathbf{x}^T)(\hat{\beta} - \beta)$ (see Fan and Li, 2001). The performances of a few other model selection alternatives are also shown for comparison: namely, the full model, the oracle model (true subset only, as if it were known in advance), and the subset chosen by significance testing on each coefficient of the full model at $\alpha \approx .05$, with coefficients either refit after selection ("Test & Refit") or not ("Naïve Test"). Tests were based on the robust (sandwich) GEE standard error estimate as suggested by Liang and Zeger (1986); predictor j was included if $|\hat{beta}_j|/\text{SE}(\hat{beta}_j) > 2$.

As we would hope, the choice of formula among (33) through (38) had little effect on performance, although in general, formulas (34) and (35) had lower rates of erroneous deletions. Similarly, the three BIC-like penalty sizes – $\log(n)$, $\log(\tilde{n})$, and $\log(N)$ – do not differ greatly in performance, although

the lighter penalty $\log(n)$ of course had the fewest erroneous deletions and fewest correct deletions. Other results were not surprising: stronger penalties generally led to more deletions, and a correct working structure usually led to better overall performance than an incorrect one. Methods with fewer erroneous deletions had somewhat more erroneous inclusions (i.e., fewer correct deletions), because of the unavoidable tradeoff between Type I and Type II error. None of the model selection methods performed nearly as well as the oracle, since this simulation uses a very difficult selection scenario.

All of these versions of AIC are in the spirit of the classic likelihood (30). If desired, still more possible generalizations could be constructed by considering a restricted likelihood (REML) function, as in Gurka (2006). However, let us instead consider extensions of the combined shrinkage and selection criteria of Section 2.2.

3.1.5 Generalizing LASSO and SCAD

Marx (2000) considered various possibilities for penalized GEE as a way of addressing collinearity in the predictors. He insightfully discussed ideas for principal components regression, Stein, ridge, or other shrinkage, and iteratively reweighted partial least squares for GEE, analogous to the corresponding classic techniques for independent linear models. However, he did not test these ideas extensively, and his focus was on regularization under an ill-specified design matrix rather than model selection.

More recently, Fu (2003) proposed a generalization of the bridge and LASSO penalties to GEE models. The usual bridge regression (L_q penalty, e.g., LASSO if $q = 1$) minimizes the penalized deviance criterion

$$-2\ell(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \mathcal{P}(\boldsymbol{\beta}) \tag{43}$$

where $\mathcal{P}(\boldsymbol{\beta}) = \sum_j p(\lambda(|\beta_j|)) = \lambda \sum_j |\beta_j|^q$, by solving the penalized generalized estimating equations (PGEE)

$$\begin{cases} s_{n1}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \dot{\mathcal{P}}_1 = 0 \\ \vdots \\ s_{nd}(\boldsymbol{\beta}, \mathbf{X}, \mathbf{y}) + \dot{\mathcal{P}}_d = 0 \end{cases} \tag{44}$$

where the s_{nj} are the GEE quasi-score functions (25) and

$$\dot{\mathcal{P}}_j = \sum_j p_\lambda(|\beta_j|) = \lambda \sum_j q|\beta_j|^{q-1} \operatorname{sgn}(|\beta_j|) \tag{45}$$

Table 2: **Comparison of AIC- and BIC-Like Criteria for GEE with AR(1) (Autoregressive) True Correlation**. For AIC and BIC, six formulas were tried; see page 22. For BIC, three sample size measures were tried: light (n), medium (Fu's $\tilde{n}$; see 49), and heavy (N). Performance measures are defined on page 23. "Ind", "Ar" and "CS" represent independence, AR(1), and compound symmetric working covariance structures.

	Mean Model Error			Correct Deletions			Wrong Deletions		
Criterion	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	1.35	0.72	1.03	0.00	0.00	0.00	0.00	0.00	0.00
AIC (33)	1.16	0.78	0.94	5.11	5.37	5.25	0.84	0.94	0.91
AIC (34)	1.16	0.65	0.94	5.11	5.32	5.11	0.84	0.65	0.84
AIC (35)	1.16	0.65	0.94	5.11	5.32	5.11	0.84	0.65	0.84
AIC (36)	1.16	0.77	0.94	5.11	5.38	5.25	0.84	0.93	0.91
AIC (37)	1.16	0.78	0.94	5.11	5.37	5.25	0.84	0.94	0.91
AIC (38)	1.16	0.77	0.94	5.11	5.38	5.25	0.84	0.93	0.91
BIC (33) (n)	1.13	0.88	0.99	5.63	5.78	5.71	1.12	1.18	1.20
BIC (34) (n)	1.13	0.65	0.98	5.63	5.77	5.63	1.12	0.90	1.13
BIC (35) (n)	1.13	0.65	0.98	5.63	5.77	5.63	1.12	0.90	1.13
BIC (36) (n)	1.13	0.88	0.99	5.63	5.78	5.71	1.12	1.18	1.20
BIC (37) (n)	1.13	0.87	0.99	5.61	5.78	5.65	1.10	1.17	1.18
BIC (38) (n)	1.13	0.87	0.99	5.61	5.78	5.67	1.10	1.17	1.18
BIC (33) (Fu $\tilde{n}$)	1.18	0.95	0.98	5.85	5.86	5.79	1.34	1.30	1.22
BIC (34) (Fu $\tilde{n}$)	1.18	0.65	0.98	5.85	5.89	5.76	1.34	0.98	1.20
BIC (35) (Fu $\tilde{n}$)	1.18	0.65	0.98	5.85	5.89	5.76	1.34	0.98	1.20
BIC (36) (Fu $\tilde{n}$)	1.18	0.95	0.98	5.85	5.86	5.79	1.34	1.30	1.22
BIC (37) (Fu $\tilde{n}$)	1.19	0.94	0.99	5.82	5.85	5.78	1.34	1.29	1.22
BIC (38) (Fu $\tilde{n}$)	1.19	0.94	0.99	5.82	5.85	5.78	1.34	1.29	1.22
BIC (33) (N)	1.18	1.00	1.04	5.85	5.91	5.87	1.34	1.36	1.35
BIC (34) (N)	1.18	0.67	1.03	5.85	5.92	5.85	1.34	1.05	1.33
BIC (35) (N)	1.18	0.67	1.03	5.85	5.92	5.85	1.34	1.05	1.33
BIC (36) (N)	1.18	1.00	1.04	5.85	5.91	5.87	1.34	1.36	1.35
BIC (37) (N)	1.19	0.97	1.04	5.82	5.90	5.86	1.34	1.34	1.34
BIC (38) (N)	1.19	0.97	1.04	5.82	5.90	5.86	1.34	1.34	1.34
Oracle Model	0.76	0.48	0.61	6.00	6.00	6.00	0.00	0.00	0.00

Table 3: **Comparison of AIC- and BIC-Like Criteria for GEE with Compound Symmetric (Exchangeable) True Correlation.** The notation here is the same as in Table 2.

	Mean Model Error			Correct Deletions			Wrong Deletions		
Criterion	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	1.49	0.97	0.90	0.00	0.00	0.00	0.00	0.00	0.00
AIC (33)	1.30	0.98	0.96	5.12	5.56	5.59	0.77	0.91	0.94
AIC (34)	1.30	0.87	0.90	5.12	5.26	4.55	0.77	0.55	0.59
AIC (35)	1.30	0.87	0.90	5.12	5.26	4.55	0.77	0.55	0.59
AIC (36)	1.30	0.96	0.96	5.12	5.57	5.59	0.77	0.90	0.93
AIC (37)	1.30	0.98	0.96	5.10	5.54	5.51	0.76	0.90	0.91
AIC (38)	1.30	0.96	0.96	5.10	5.53	5.51	0.76	0.89	0.91
BIC (33) (n)	1.30	1.07	1.05	5.73	5.83	5.87	1.20	1.21	1.21
BIC (34) (n)	1.30	0.88	0.97	5.73	5.73	5.72	1.20	0.87	1.01
BIC (35) (n)	1.30	0.88	0.97	5.73	5.73	5.72	1.20	0.87	1.01
BIC (36) (n)	1.30	1.07	1.05	5.73	5.83	5.87	1.20	1.21	1.21
BIC (37) (n)	1.29	1.07	1.05	5.73	5.83	5.86	1.19	1.20	1.20
BIC (38) (n)	1.29	1.07	1.05	5.73	5.83	5.86	1.19	1.20	1.20
BIC (33) (Fu $\tilde{n}$)	1.38	1.12	1.06	5.83	5.88	5.87	1.42	1.31	1.23
BIC (34) (Fu $\tilde{n}$)	1.38	0.90	0.98	5.83	5.84	5.78	1.42	1.01	1.08
BIC (35) (Fu $\tilde{n}$)	1.38	0.90	0.98	5.83	5.84	5.78	1.42	1.01	1.08
BIC (36) (Fu $\tilde{n}$)	1.38	1.12	1.06	5.83	5.88	5.87	1.42	1.31	1.23
BIC (37) (Fu $\tilde{n}$)	1.37	1.10	1.06	5.83	5.86	5.87	1.39	1.29	1.22
BIC (38) (Fu $\tilde{n}$)	1.37	1.11	1.06	5.83	5.86	5.87	1.39	1.29	1.22
BIC (33) (N)	1.38	1.17	1.18	5.83	5.93	5.92	1.42	1.43	1.46
BIC (34) (N)	1.38	0.94	1.07	5.83	5.92	5.89	1.42	1.12	1.25
BIC (35) (N)	1.38	0.94	1.07	5.83	5.92	5.89	1.42	1.12	1.25
BIC (36) (N)	1.38	1.17	1.18	5.83	5.93	5.92	1.42	1.43	1.46
BIC (37) (N)	1.37	1.16	1.17	5.83	5.92	5.92	1.39	1.42	1.45
BIC (38) (N)	1.37	1.16	1.17	5.83	5.92	5.92	1.39	1.42	1.45
Oracle Model	0.88	0.68	0.66	6.00	6.00	6.00	0.00	0.00	0.00

Thus we are solving

$$\sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i(\boldsymbol{\beta})) + \dot{\mathcal{P}} = 0 \tag{46}$$

which, if we treat $\mathbf{V}^{-1}$ as fixed, this would be the same as minimizing a penalized generalized least squares (GLS) criterion

$$\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \mathcal{P}(\boldsymbol{\beta}), \tag{47}$$

Fu suggests solving (46) either by adjusting the iteratively reweighted least squares method for the penalty function (this is equivalent to the LQA algorithm of Section 2.2.4) or, in the case of LASSO, accomodating his specialized "shooting" method (see Fu, 1998).

Fu also considers how to choose λ, recommending an adaptation of the GCV (21). The classical RSS is generalized to the weighted deviance

$$\text{WDev} = \sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{R}_i^{-1} \mathbf{r}_i \tag{48}$$

to take into account correlations and to allow non-Gaussian responses. Here $\mathbf{R}$ is the working correlation matrix, and the $\mathbf{r}_i$ are deviance residuals (see McCullagh and Nelder, 1991; Agresti, 2002), although they could also be reasonably replaced by the Pearson residuals $\mathbf{A}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$ for simplicity; in either case, for a normal linear model both deviance and Pearson residuals reduce to $\mathbf{y} - \boldsymbol{\mu}$. For the degrees of freedom Fu uses df_s where $s = \|\boldsymbol{\beta}(\lambda)\| \,/\, \|\boldsymbol{\beta}\|$ and $\boldsymbol{\beta}$ is the vector of unpenalized estimates for the selected coefficients. As we saw with BIC, n in the GCV formula is now ambiguous. Fu also suggests a compromise between counting subjects and counting all observations:

$$\tilde{n} = \sum_{i=1}^{n} \frac{J_i^2}{\sum_{j=1}^{n} \sum_{k=1}^{n} \mathrm{R}_{jk}} \tag{49}$$

Thus $n \leq \tilde{n} \leq nJ$, and $\tilde{n}$ becomes smaller as observations become more correlated. We can now choose λ by minimizing

$$\text{QGCV} = \tilde{n}^{-1}\text{WDev} / \left(1 - \mathrm{df}_F/\tilde{n}\right) \tag{50}$$

over a grid; Fu suggests a range of about $[1, 100]$.

Dziak and Li (2006) recently proposed SCAD for GEE modeling and showed that this may provide better estimation and selection performance than the LASSO. They replace the bridge penalty in (43) and (46) with the

SCAD penalty. To use the SCAD effectively here, it is best to rescale the penalty and reexpress (46) as

$$\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \mu(\boldsymbol{\beta})) + N\dot{\mathcal{P}} = 0, \tag{51}$$

i.e., solve for the minimum of

$$(\mathbf{y} - \boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y} - \mu) + N\mathcal{P}(\boldsymbol{\beta}), \tag{52}$$

since $(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{V}^{-1}(\mathbf{y}-\mu) = O_p(N)$. Rescaling the penalty does not matter for the LASSO as long as we also rescale the tuning parameter accordingly, since the LASSO penalty is proportional to λ, so $\mathcal{P}(\boldsymbol{\beta}, \lambda) = N\mathcal{P}(\boldsymbol{\beta}, \lambda/N)$. However, this relationship does not hold for the SCAD, since the relative sizes of λ, β_j and a determine the shapes of functions (9) and (10) (see Sections 3.4 and **??**).

3.2 Measuring Model Complexity for Correlated Data and Non-Classical Penalties

The smoothing parameter λ for LASSO, SCAD, and similar techniques is chosen by optimizing some criterion which balances goodness of fit and model complexity, generally an adaptation of a classical criterion such as C_p, GCV, or BIC. Implementing these tuning parameter selection criteria here is complicated for two reasons. First, because we are combining shrinkage and selection, it is not clear how to define an effective degrees of freedom. Second, because we have clustered data, it is not clear how to define goodness of fit or sample size.

For goodness of fit, we could use any of the criteria in Section 3.1.3, e.g., $N\log(\text{WSS}/N) + c_n df$. For degrees of freedom, we could use any of the measures in Section 2.2.6; note that the effective df_T for linear models here generalizes to

$$df_T \equiv \text{tr}(\mathbf{P}_x) \quad \text{where} \quad \mathbf{P}_x = \tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\mathbf{R}_{\cdot}^{-1}\tilde{\mathbf{X}} + N\Delta\right)^{-1}\tilde{\mathbf{X}}^T\mathbf{R}_{\cdot}^{-1}, \tag{53}$$

where $\mathbf{R}_{\cdot}$ is the block-diagonal matrix of all $\mathbf{R}_i$'s. We might also use the df estimates derived by Pan or Cantoni for their GEE selection criteria. Thus, as in Section 3.1.2, many tuning selectors are possible. Some of these possibilities are tested empirically in Section 3.5.

3.3 Asymptotic Properties of Penalized GEE

In this subsection I prove the $\sqrt{n}$-consistency, model selection consistency (in the BIC sense), and asymptotic normality of SCAD-penalized GEE, as well as $\sqrt{n}$ consistency and asymptotic normality for LASSO-penalized GEE. Fu (2003) had earlier given a normality proof for L_q penalties with $q > 1$, including LASSO as a limit case, but the approach used here, following Fan and Li (2001), applies much more generally.

Let $\mathbf{K}_n = \frac{1}{n}\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$ and $\mathbf{s}_n(\beta) = \sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$. Note that $\mathbf{V}_i$ here is only the working covariance, and is not assumed to be the same as the true $\mathrm{Cov}(\mathbf{y}_i|\boldsymbol{\beta}_i)$; however, for convenience I do treat $\mathbf{D}_i$ as known. For example, in the Gaussian case $\mathbf{D}_i = \mathbf{X}_i$; in general with the canonical link it is $\mathbf{D}_i = \mathbf{A}_i \mathbf{X}_i$.

Suppose we wish to choose an estimate $\hat{\beta}$ by minimizing the penalized loss criterion

$$Q_n^{\mathcal{P}}(\beta) = \frac{1}{2n}\mathbf{s}_n^T \mathbf{K}_n^{-1} \mathbf{s}_n + k_n \mathcal{P}(\beta). \tag{54}$$

where k_n is a $O(n)$ sequence bounded above zero; e.g., $k_n = 1$, n, or N. The scaling factor k_n is included for convenience so that the results can be easily interpreted either in the parameterization of 44 or 51; however, we will use $k_n = N$ for convenience and for consistency with previous literature. Treating $\mathbf{K}_n$ as constant and noting $n^{-1}\partial \mathbf{s}_n(\beta)/\partial\beta = \mathbf{K}_n$, minimizing (54) is just solving the penalized GEE equation $\mathbf{s}_n + k_n \dot{\mathcal{P}}(\beta) = 0$. This observation is convenient because we can characterize the resulting $\hat{\beta}$ as a penalized generalized least squares solution, and therefore can easily adapt the proofs in Fan and Li (2001).

Let $\boldsymbol{\beta}_0$ be the fixed true value of $\boldsymbol{\beta}$ and let $n \to \infty$ while the J_i are bounded uniformly between 1 and ∞. Then under some regularity conditions, we have the following theorem generalizing Theorem 1 in Fan and Li (2001). Proofs are given at the end of the section.

Theorem 3.1 *($\sqrt{n}$-consistent estimation) There exists a sequence $\hat{\boldsymbol{\beta}}_n$ of solutions of (51) such that $\left\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| = O_p(n^{-1/2})$.*

For "consistent" model selection, which is in turn the basis for the oracle property, two properties are required: sparsity (deleting all of the coefficients which should be deleted, with probability approaching one), and sensitivity (retaining all of the coefficients which should be retained, $wp \to 1$). To establish these, partition $\boldsymbol{\beta}_0$ into active (nonzero) and inactive (zero) coefficients

as follows: let $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ and $\mathcal{N} = \{j : \beta_{0j} \neq 0\}$. Let s be the cardinality of $\mathcal{A}$ (note that $0 \leq s \leq d$). Denote the active and inactive coefficients themselves as $\boldsymbol{\beta}_{\mathcal{A}}$ and $\boldsymbol{\beta}_{\mathcal{N}}$ respectively. Then we have the following lemma.

Lemma 3.1 *(Sensitivity) The active coefficients are included in the model with probability approaching one, i.e.,* $\Pr(\exists\ j \in \mathcal{A} : \hat{\beta}_j = 0) = o(1)$.

Now suppose further that $\forall j$ there exist sequences λ_{nj} suc h that

$$\liminf_{n\to\infty} \left(\liminf_{\beta_j \to 0^+} \lambda_{nj}^{-1} p'_{nj}(\beta_j) \right) > 0 \tag{55}$$

and

$$k_n n^{-1/2} \lambda_{nj} \to \infty \tag{56}$$

The λ_{nj} coincide with the tuning parameters for the SCAD and LASSO penalties. However, for the LASSO, (56) is incompatible with Condition 3.5. This is why the oracle property is possible for SCAD but not for LASSO (Fan and Li, 2001).

Lemma 3.2 *(Sparsity) There exists a sequence $\boldsymbol{\beta}_n$ of solutions of (51) such that $\boldsymbol{\beta}_{\mathcal{A}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_{0\mathcal{A}}$, and that wp $\to 1$, $\boldsymbol{\beta}_{\mathcal{N}} = \boldsymbol{\beta}_{0\mathcal{N}} = \mathbf{0}_{d_n - s_n}$.*

Thus, the appropriately penalized SCAD estimator is a consistent model selector, so its asymptotic distribution is normal with bias and variance not much larger than they would be in the absence of the nuisance $\boldsymbol{\beta}_{\mathcal{N}}$. Thus, we have an asymptotic "oracle property" as described in Fan and Li (2001) and Li et al. (2006). This does not hold for the LASSO, since by making λ high enough to achieve appropriate sparsity, we would lose $\sqrt{n}$-consistency. We state the oracle property as follows, analogously to Theorem 2 in Fan and Li (2001) (see also Harris and Mátyás (1999), pp. 19-20, and Liang and Zeger (1986)).

Let the vector $\dot{\mathcal{P}}_{n\mathcal{A}}$ and the matrix $\ddot{\mathcal{P}}_{n\mathcal{A}}$ be the first and second derivatives of $\mathcal{P}$ in $\boldsymbol{\beta}_{\mathcal{A}}$. Alternatively, to be more careful, one could define the $\dot{\mathcal{P}}_{nj}$ and $\ddot{\mathcal{P}}_{nj}$ as the appropriate derivatives from the right, to allow for the cases where some $\beta_{\mathcal{A}}$ are mistakenly estimated as zero, so that $\mathcal{P}_{n\mathcal{A}}$ itself is not differentiable. $\dot{\mathcal{P}}$ and $\ddot{\mathcal{P}}$ will always be proper derivatives when evaluated at the true $\boldsymbol{\beta}_{n\mathcal{A}}$, since by definition the $\boldsymbol{\beta}_{n\mathcal{A}}$ are all nonzero. The nondifferentiability at zero for $q \leq 1$ was what kept Fu from addressing L_q penalties more generally.

We assume $\ddot{\mathcal{P}}$ is a diagonal matrix, i.e., that the penalty applied to one coefficient does not depend on the values of other coefficients. It might be

interesting and useful to relax this in the future, by letting the penalties for certain coefficients depend on the values of other coefficients. This might be done to penalize multicollinearity somehow, or to encourage hierarchical structure among the predictors, e.g., let the penalty for an interaction term depend on whether, and to what extent, the parent terms are included in the model.

The following theorem is analogous to Theorem 2 of Fan and Li (2001).

Theorem 3.2 *(Asymptotic Normality) Let* $\mathbf{\Sigma}_i$ *be the unknown true* $Cov(y_i)$ *and let*

$\mathbf{C}_n = \sum_{i=1}^n \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{\Sigma}_i \mathbf{V}_i^{-1} \mathbf{D}_i$. *Let*

$$\begin{aligned} &\mathbf{H}_{n0\mathcal{A}} = \mathbf{K}_{n0\mathcal{A}}^T, \\ &\mathbf{K}_{n0}^{-1}\mathbf{K}_{n0\mathcal{A}}, \text{ and} \\ &\mathbf{S}_{n0\mathcal{A}} = \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{C}_{n0} \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}}. \end{aligned} \tag{57}$$

Then

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\right) + \mathbf{b}_n \xrightarrow{L} \mathrm{N}\left(\mathbf{0}, \boldsymbol{Phi}\right) \tag{58}$$

where $\mathbf{\Phi}$ *is the limit in probability of*

$$\mathbf{\Phi}_n = \frac{1}{n}\left(\mathbf{H}_{n\mathcal{A}} + \frac{k_n}{n}\mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0)\right)^{-1} (\mathbf{S}_{n\mathcal{A}}) \left(\mathbf{H}_{n\mathcal{A}} + \frac{k_n}{n}\mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0)\right)^{-1} \tag{59}$$

where $\mathbf{H}_{n\mathcal{A}} = \mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}$, $\mathbf{S}_{n\mathcal{A}} = \mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{C}_n \mathbf{K}_n^{-1} \mathbf{K}_{n\mathcal{A}}$, $\mathbf{K}_{n\mathcal{A}} = \frac{1}{n}\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_{i\mathcal{A}}$, *and* $\mathbf{D}_{i\mathcal{A}} = \partial_i / \partial \boldsymbol{\beta}_{\mathcal{A}}$, *and where the bias term* $\mathbf{b}_n = k_n n^{-1/2} \mathcal{P}_{n\mathcal{A}}^*(\boldsymbol{\beta}_0) = O_p(1)$, *so that* $n^{-1/2}\mathbf{b}_n$ *vanishes asymptotically.*

Proofs are given at the end of the section. Notice that if we substitute a robust (empirical) estimator of $\mathrm{Cov}(y_i)$ for $\mathbf{\Sigma}_i$, and the full model is true, then the PGEE sandwich covariance estimate (59) reduces to the robust "sandwich" covariance estimator of Liang and Zeger (1986). (59) is also very similar to the sandwich variance estimate given in Fan and Li (2001) for SCAD-penalized likelihood estimators.

Let us examine the asymptotic variance more closely. By Condition 3.6, the penalty term in (59) is $o_p(1)$ and dominated by $\mathbf{H}_{n\mathcal{A}}$, so we could also write the asymptotic covariance more coarsely as $n^{-1}\mathbf{H}_{n\mathcal{A}}^{-1}\mathbf{S}_{n\mathcal{A}}\mathbf{H}_{n\mathcal{A}}^{-1}$. If the covariance is correctly specified, $\mathbf{C}_n = \mathbf{K}_n$, so the covariance then becomes $\mathbf{H}_{n\mathcal{A}}^{-1}$. This corresponds to the superefficiency or "oracle property" of Fan and

Li (2001). To see why, suppose for simplicity that $\sigma^2 = 1$. Then $\mathbf{K}_{n\mathcal{A}} = \mathbf{X}^T\mathbf{X}_{\mathcal{A}}^T$ so $\mathbf{H}_{n\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}$, and so the asymptotic variance for the SCAD estimate is the same as that which we would have if we knew in advance the correct subset and did not have to select a model or deal with the nuisance null coefficients. It is not yet clear to what extent this oracle property still holds in the GEE and QIF models considered in this thesis, since the presence of the nuisance coefficients may still affect covariance parameter estimation.

Last, notice the unusual multiplier for the penalty, k_n/n. For balanced data and the recommended $k_n = N$ this will equal J (which was 1 for Fan and Li (2001)). This is not intuitive at first but does make sense. Imagine a simple situation: a linear model with working independence and an L_2 penalty (so that no predictors are excluded). Then $\mathbf{H}_{n\mathcal{A}} = \frac{1}{n}\sum_i \mathbf{X}_i^T\mathbf{X}_i = \frac{1}{n}\sum_i\sum_J \mathbf{x}_i^T\mathbf{x}_i$, which approaches a constant for fixed J and growing n but approaches infinity for fixed n and growing J. Therefore, if the penalty is to have the same meaning for different cluster sizes, it must be able to keep up with J.

3.4 Scaling the Penalty in Penalized GEE

We minimize (54) by solving $\mathbf{s}_n(\beta) + k_n\dot{\mathcal{P}}(\beta) = \mathbf{0}$. It is not clear at first how to choose k_n for GEE, because of the usual ambiguity of the sample size ($n \neq N$). However, this question can be addressed by comparison with previous results, and N rather than n is recommended.

The choice of k_n here is of interest because it determines the meaning of λ. For SCAD, it also determines the shape of the penalty function; if we pick the wrong k_n then the estimator will not behave as we expect (see p. 3.1.5). To see why this is so, consider the penalized least squares case with N independent observations, d orthogonal predictors, all x scaled and centered (so that $\mathbf{X}^T\mathbf{X} = N\mathbf{I}$), and y centered (so that we do not need to consider an intercept). . Suppose we wish to minimize

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + N\sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{60}$$

which was the form of penalized least squares used by Fan and Li (2001) and Li et al. (2006). (Section 2 of Fan and Li (2001) appeared to use $k_n = 1$ but in fact they were still implicitly using $k_n = N$, since they assumed $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ and hence rescaled the observations.) Denote the least-squares (unpenalized)

estimate of $\boldsymbol{\beta}$ as $\mathbf{z}$. Since the predictors are orthogonal, the $\hat{\beta}_j$ will have the same signs as the corresponding z_j, so for convenience assume z_j is positive. Thus the penalized criteria will be minimized at

$$\hat{\beta}_j = \begin{cases} 0 & \text{if } z_j \leq p_0 \\ z_j - p'_\lambda(z_j) & \text{if } z_j > p_0 \end{cases}$$

where the threshold p_0 equals $min_{\beta \geq 0}(\beta + p'_\lambda(\beta))$ (see Antoniadis and Fan, 2001; Fan and Li, 2001; Li et al., 2006) Suppose now we replace the N in (60) with k_n, i.e., replace the penalty $p_\lambda(\cdot)$ in (60) with $c_\lambda(\cdot) \equiv N^{-1}k_n p_\lambda^{(old)}(\cdot)$. What happens to the resulting penalized estimate of β?

Consider first the LASSO penalty $p_\lambda(|\beta_j|) = \lambda_n |\beta_j|$. Then for positive β_j, $c'_\lambda(\beta_j) \equiv N^{-1}k_n\lambda_n$ and so $p_0^{new} \equiv min_{\beta \geq 0}(\beta + c'_\lambda(\beta_j)) = N^{-1}k_n\lambda_n$. Thus

$$\hat{\beta}_j = \begin{cases} 0 & \text{if } z_j \leq N^{-1}k_n\lambda_n \\ z_j - N^{-1}k_n\lambda_n & \text{if } z_j > N^{-1}k_n\lambda_n \end{cases}$$

i.e., we have a soft thresholding rule (see Donoho and Johnstone, 1994; Fan and Li, 2001). Comparing this with the corresponding entry in Table 1, it is clear that using k_n instead of N simply rescales λ to $N^{-1}k_n\lambda_n$. Thus if we express the criterion as $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^d p_\lambda(|\beta_j|)$, so that $k_n = 2$, we get the same estimate for a given λ^* as we would get using $k_n = N$ and $\lambda = 2\lambda^*/N$. If we use a smaller k_n then we get much less shrinkage and less likelihood of deletion for a given λ, but this does not matter because we can rescale λ_n accordingly to get the desired behavior. Thus it is not a problem that, say, Fu (2003) recommended a λ in the range of 1 to 100, i.e., $O(\sqrt{N})$, in contrast to Fan and Li (2001) who recommended $O(N^{-1/2})$; the difference is only in the parameterization. We would even not see the difference in practice, because λ would be chosen in an adaptive, data-driven way (e.g., cross-validation, C_p, GCV, BIC) rather than set to a specific value. However, among these logically equivalent parameterizations, (60) with $k_n = N$ seems preferable for interpretation, because λ_n is then the threshold parameter in the orthogonal case, and otherwise λ has no direct interpretation.

For SCAD the situation is more complicated because of the additional parameter a, which does not respond to scaling in the same way. Now, if we choose the wrong k_n, there may be no way to rescale λ_n to recover the behavior we desire. This complication is not specific to SCAD, but could be encountered with any penalty function for which the tuning parameter is other than a multiplier, i.e., for which we do not have $p_{\lambda_1}(\cdot) = \frac{\lambda_1}{\lambda_2} p_{\lambda_2}(\cdot) \, \forall \lambda_1, \lambda_2$, and

in practice we would only have the latter with the L_q penalties. However, under the wrong k_n the SCAD estimate may no longer be a continuous function of $\mathbf{z}$ as it was in Antoniadis and Fan (2001), Fan and Li (2001) and Li et al. (2006), so we need to be sure of choosing k_n correctly, so that SCAD behaves in the same way as in those previous papers. It seems at first that $k_n = n$ is more reasonable, since asymptotic consistency and normality requires $n \to \infty$ and not just $N \to \infty$ (trivially, imagine a situation in which each subject had a large number J of completely identical observations). However, what is important is that the objective function (e.g., RSS) grows both in n and J, regardless of the degree of within-subject correlation, so k_n must be growing both in n and J in order to keep pace; otherwise the selected model will become less parsimonious as J grows, even under working independence.

3.5 Empirical Comparison of Existing Approaches for GEE Model Selection

It would be helpful to empirically compare the performance of these methods for GEE model selection. Therefore, several simulations were performed under different conditions.

3.5.1 Initial Simulation of Penalized GEE in the Gaussian Case

Therefore, 200 simulations were performed, proceeding almost exactly as in Section 3.1.3. The same true model and parameters as in 3.1.3 were used here, except for using a more reasonable σ of 3 instead of 6 for the errors. [1] The performances of eight different discrete subset-selection methods, ten different tuning parameter selectors for the L_1 penalty, and ten different tuning parameter selectors for the SCAD penalty, representing the possibilities described in Sections 3.1.5, 3.2, and 3.3, were compared.

The results are shown in Tables 4, 6, 5, and 7. Results are substantively

[1] Note that $\boldsymbol{\beta}^T E(\mathbf{y}\mathbf{y}^T)\boldsymbol{\beta} = 31.05$ here, so the marginal variance of y was about 67; compared to this the regression coefficients 3, 2, 1, 0.5 were quite small. In the current subsection $Var(y|\mathbf{y}\boldsymbol{\beta}) = 9$ so $Var(y) \approx 40$. The rather large variances here are used to keep the selection task from becoming too easy in light of the large number of observations; recall that the standard error of regression coefficients will be on the order of $\sigma/\sqrt{N}$, so with $N = 350$ and $\sigma = 3$ we get standard errors around .2. If the standard error were much smaller in relation to the smallest regression coefficient, erroneous deletions would become almost nonexistent, and the performance of selection methods would then be determined almost entirely by their estimation bias and Type I error rate, thus unfairly favoring BIC.

similar for true AR(1) correlation (Tables 4 and 5) and for true compound symmetric correlation (Tables 6 and 7) so they are discussed together below.

As shown in Tables 4 and 6, the methods are generally similar in terms of model error, but the LASSO-type (L_1-penalized) methods delete considerably fewer variables and thus have fewer correct deletions and fewer wrong deletions. In general, very good performance was obtained by SCAD with a lighter ($\log(n)$) BIC-like penalty selector. Usually, performance did not greatly depend on working covariance structure, despite the fairly strong ($\rho = .7$) within-subject correlation parameter. Working indepenedence tended to be associated with higher rates of deletions, especially of wrong deletions, possibly because it led to a poorer variance estimate.

Details of the formulas for the selection criteria did not greatly matter. In terms of overall model error, Pan's quasi-AIC and Cantoni's generalized C_p give very similar results to those which would be obtained by naïvely using the corresponding classical formulas ignoring correlation, although they do enjoy a lower rate of erroneous deletions. Similarly, for BIC it did not much matter whether we used form (33) or (34), or whether we used $\log(n)$ or $\log(N)$ in the penalty. For the shrinkage-based methods, using df_T was generally not much different than using df_s, although df_T tended to choose a larger model (smaller λ) than df_s since for a given λ, $df_s \leq df_T$. Fu's quasi-GCV selects smaller models than a naïve GCV because the former uses Fu's effective sample size and the latter uses N.

The tests, which had nominal α of .05, would be expected to have an average of about $6 \times .95 \approx 5.7$ correct deletions per simulation; they had almost that many. For testing, power seems to be lowest (there were more wrong deletions) when working independence is used, and highest (there were fewest wrong deletions) when the correct covariance structure is used.

Tables 5 and 7 divide the results of each of the simulations into four possible decisions: the correct subset, an overfit model (at least one erroneous inclusion but no erroneous deletion, i.e., fewer than six correct zeroes but no incorrect zeroes), an underfit model (six correct zeroes but at least one incorrect zero), or a misfit model (both erroneous inclusions and erroneous deletions).

Heavier penalties (e.g., $\log(N)$ rather than $\log(n)$, df_s rather than df_T) lead to fewer overfit models but more underfit models. In practice, whether an overfit model is worse than an underfit, or vice versa, is unclear and depends on the researcher's goals. In the classical hypothesis testing framework,

false inclusions (Type I errors) are considered much worse than false deletions (Type II errors). For a strictly predictive model, a false inclusion is not very serious, as long as its coefficient estimate is small; it may simply add noise. A false deletion may also create a confounding variable and render the model, in some sense, invalid; while a false inclusion generally will not do so. In a decision-based context, the relative seriousness of false exclusion or inclusion depends on the expected practical outcome, e.g., if we are modeling mortality in terms of various possibly carcinogenic pollutants, Type II error is more hazardous to public health than Type I. However, in any case, misfit models are clearly undesirable, so it is important to note that the LASSO and SCAD criteria have lower misfitting rates than the best subset criteria. The reason for this is unclear, but it may be related to the intuition that shrinkage methods are better able to deal with multicollinearity than classic selection methods (see Fu, 2003; Babyak, 2004).

Unfortunately, the rate of correct model identification is not very high for any of the methods, since a correct model requires a correct decision for all ten predictors. However, in general the testing methods, the naïve BIC criteria, and SCAD with a BIC selection criterion do better than the other methods. The L_1 methods almost always overfit rather than choosing the correct model; this is not surprising in light of the findings of Fan and Li (2001) and Leng et al. (2004).

3.5.2 Gaussian Simulation with Unbalanced Cluster Sizes

Similar simulations were also performed for unbalanced cluster sizes, in which there were still 50 subjects but now each subject i had a random number of observations J_i between 5 and 12, inclusive. The results of these are shown in somewhat condensed form in Tables 8 and 9. Following Fan and Li (2001), model error in Tables 8 and 9 has been reexpressed as mean relative model error, i.e., the mean ratio of the ME obtained by a given combination of model selection and working covariance to that of the working-independence full model, i.e.,

$$\text{RME} = \frac{\text{ME}(\hat{\beta}(\text{selected model}))}{\text{ME}(\hat{\beta}(\text{OLS}))} \tag{61}$$

This is done to make results easier to interpret, since, say, an RME of .8 means that a method led to estimates with about 80% as much error as those which would be obtained from naïvely using ordinary least squares.

As the results in Tables 8 and 9 are substantively similar to the previous

Table 4: **Performance Measures for Various Selection and Tuning Criteria for GEE with Autoregressive (AR-1) True Correlation**. Performance measures are the same as those in 2, but here we also try LASSO and SCAD penalties.

	Mean Model Error			Correct Deletions			Wrong Deletions		
	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	0.36	0.21	0.25	0.00	0.00	0.00	0.00	0.00	0.00
Best Subset according to...									
Naïve AIC (33)	0.29	0.21	0.20	5.46	5.53	5.49	0.40	0.39	0.31
Naïve C_p (37)	0.29	0.21	0.20	5.46	5.53	5.49	0.40	0.39	0.31
Pan AIC (28)	0.30	0.20	0.20	5.34	4.96	5.13	0.41	0.26	0.23
Cantoni C_p (27)	0.29	0.20	0.20	5.46	5.12	5.16	0.40	0.28	0.24
BIC (33) (n)	0.29	0.24	0.21	5.66	5.85	5.80	0.52	0.53	0.44
BIC (33) (N)	0.29	0.26	0.22	5.83	5.95	5.91	0.63	0.69	0.62
BIC (34) (n)	0.29	0.19	0.20	5.66	5.81	5.75	0.52	0.26	0.39
BIC (34) (N)	0.29	0.20	0.22	5.83	5.92	5.91	0.63	0.39	0.58
LASSO with λ chosen by...									
Naïve GCV (21)	0.28	0.23	0.26	2.98	1.73	1.31	0.05	0.01	0.01
Fu GCV (50)	0.29	0.28	0.34	3.01	2.49	2.50	0.06	0.01	0.01
BIC (33) (n, df_T)	0.29	0.27	0.30	3.60	2.38	1.99	0.07	0.01	0.01
BIC (33) (n, df_s)	0.29	0.26	0.29	3.74	2.28	2.00	0.07	0.01	0.01
BIC (34) (n, df_T)	0.29	0.26	0.29	3.60	2.21	1.91	0.07	0.01	0.01
BIC (34) (n, df_s)	0.29	0.24	0.28	3.74	2.07	1.82	0.07	0.01	0.01
BIC (34) (N, df_T)	0.32	0.30	0.34	3.85	2.61	2.28	0.08	0.01	0.01
BIC (34) (N, df_s)	0.30	0.27	0.31	3.89	2.44	2.26	0.08	0.01	0.01
BIC (33) (N, df_T)	0.32	0.30	0.34	3.85	2.61	2.28	0.08	0.01	0.01
BIC (33) (N, df_s)	0.30	0.27	0.31	3.89	2.44	2.26	0.08	0.01	0.01
SCAD with λ chosen by...									
Naïve GCV (21)	0.29	0.18	0.19	4.35	3.78	3.87	0.20	0.02	0.08
Fu GCV (50)	0.31	0.20	0.23	4.92	5.68	5.82	0.35	0.21	0.40
BIC (34) (n, df_T)	0.27	0.18	0.19	5.20	4.74	4.96	0.33	0.05	0.14
BIC (34) (n, df_s)	0.29	0.19	0.20	5.50	5.51	5.53	0.50	0.15	0.30
BIC (33) (n, df_T)	0.27	0.22	0.20	5.20	5.64	5.47	0.33	0.16	0.23
BIC (33) (n, df_s)	0.29	0.21	0.20	5.50	5.72	5.62	0.50	0.25	0.31
BIC (33) (N, df_T)	0.28	0.19	0.21	5.54	5.30	5.51	0.45	0.10	0.23
BIC (33) (N, df_s)	0.30	0.20	0.21	5.67	5.71	5.73	0.59	0.22	0.34
BIC (34) (N, df_T)	0.28	0.19	0.21	5.54	5.30	5.51	0.45	0.10	0.23
BIC (34) (N, df_s)	0.30	0.20	0.21	5.67	5.71	5.73	0.59	0.22	0.34
Oracle	0.19	0.15	0.14	6.00	6.00	6.00	0.00	0.00	0.00

Table 5: **Selected Models for Various Selection and Tuning Criteria for GEE with Autoregressive (AR-1) True Correlation.** The columns labeled C, O, U, and M show the proportion of trials in which the correct model, an overfit model, an underfit model, or a misfit model was selected by each method; see page 35 for definitions of these terms.

	Working Indep				Working AR-1				Working CS			
	C	O	U	M	C	O	U	M	C	O	U	M
Full Model	.00	1	.00	.00	.00	1	.00	.00	.00	1	.00	.00
Best Subset according to...												
Naïve AIC (33)	.35	.26	.21	.19	.44	.18	.20	.19	.43	.27	.14	.18
Naïve C_p (37)	.35	.26	.21	.19	.44	.18	.20	.19	.43	.27	.14	.18
Pan AIC (28)	.33	.27	.20	.21	.29	.46	.07	.19	.33	.45	.06	.17
Cantoni C_p (27)	.35	.26	.21	.19	.33	.41	.08	.19	.32	.45	.06	.18
BIC (33) (n)	.36	.13	.35	.17	.45	.04	.41	.11	.50	.07	.32	.12
BIC (33) (N)	.35	.03	.50	.13	.33	.00	.63	.05	.37	.02	.54	.08
BIC (34) (n)	.36	.13	.35	.17	.65	.10	.19	.07	.53	.09	.26	.13
BIC (34) (N)	.35	.03	.50	.13	.60	.02	.32	.07	.41	.02	.50	.08
LASSO with λ chosen by...												
Naïve GCV (21)	.01	.94	.00	.05	.00	1	.00	.01	.00	1	.00	.01
Fu GCV (50)	.05	.90	.00	.06	.03	.97	.00	.01	.02	.98	.00	.01
BIC (34) (n, df_T)	.05	.88	.00	.07	.00	1	.00	.01	.00	1	.00	.01
BIC (34) (n, df_s)	.09	.84	.00	.07	.02	.98	.00	.01	.01	.99	.00	.01
BIC (33) (n, df_T)	.05	.88	.00	.07	.01	.99	.00	.01	.00	1	.00	.01
BIC (33) (n, df_s)	.09	.84	.00	.07	.03	.97	.00	.01	.01	.99	.00	.01
BIC (34) (N, df_T)	.08	.85	.00	.08	.01	.99	.00	.01	.00	1	.00	.01
BIC (34) (N, df_s)	.10	.82	.01	.08	.03	.97	.00	.01	.02	.98	.00	.01
BIC (33) (N, df_T)	.08	.85	.00	.08	.01	.99	.00	.01	.00	1	.00	.01
BIC (33) (N, df_s)	.10	.82	.01	.08	.03	.97	.00	.01	.02	.98	.00	.01
SCAD with λ chosen by...												
Naïve GCV (21)	.18	.62	.02	.19	.04	.94	.01	.01	.05	.88	.01	.07
Fu GCV (50)	.25	.41	.23	.12	.58	.22	.19	.03	.50	.12	.35	.05
BIC (34) (n, df_T)	.32	.36	.17	.16	.22	.74	.02	.03	.30	.57	.06	.09
BIC (34) (n, df_s)	.29	.22	.40	.10	.55	.30	.13	.02	.44	.27	.25	.06
BIC (33) (n, df_T)	.32	.36	.17	.16	.59	.25	.13	.03	.44	.34	.17	.06
BIC (33) (n, df_s)	.29	.22	.40	.10	.58	.18	.21	.04	.47	.23	.26	.06
BIC (34) (N, df_T)	.35	.21	.33	.12	.40	.51	.05	.05	.49	.29	.14	.09
BIC (34) (N, df_s)	.30	.13	.48	.10	.58	.21	.19	.03	.50	.17	.29	.05
BIC (33) (N, df_T)	.35	.21	.33	.12	.40	.51	.05	.05	.49	.29	.14	.09
BIC (33) (N, df_s)	.30	.13	.48	.10	.58	.21	.19	.03	.50	.17	.29	.05
Oracle	1	.00	.00	.00	1	.00	.00	.00	1	.00	.00	.00

Table 6: **Performance of Various Selection and Tuning Criteria for GEE with Exchangeable (Compound Symmetric) True Correlation.** Performance measures are the same as those in 4.

	Mean Model Error			Correct Deletions			Wrong Deletions		
	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	0.29	0.26	0.21	0.00	0.00	0.00	0.00	0.00	0.00
Best Subset according to...									
Naïve AIC (33)	0.22	0.25	0.21	5.46	5.38	5.46	0.35	0.38	0.39
Naïve C_p (37)	0.22	0.25	0.21	5.46	5.38	5.46	0.36	0.38	0.39
Pan AIC (28)	0.23	0.24	0.20	5.35	4.82	4.72	0.35	0.24	0.24
Cantoni C_p (27)	0.22	0.24	0.20	5.46	4.98	4.86	0.36	0.28	0.25
BIC (33) (n)	0.22	0.27	0.22	5.69	5.78	5.85	0.43	0.54	0.48
BIC (33) (N)	0.22	0.29	0.24	5.87	5.94	5.97	0.59	0.69	0.64
BIC (34) (n)	0.22	0.23	0.21	5.69	5.76	5.76	0.43	0.26	0.39
BIC (34) (N)	0.22	0.24	0.22	5.87	5.91	5.92	0.59	0.40	0.53
LASSO with λ chosen by...									
Naïve GCV (21)	0.22	0.28	0.22	3.16	1.40	0.78	0.04	0.00	0.00
Fu GCV (50)	0.22	0.35	0.29	3.22	2.32	1.32	0.04	0.00	0.00
BIC (34) (n, df_T)	0.23	0.31	0.25	3.78	1.97	1.01	0.05	0.00	0.00
BIC (34) (n, df_s)	0.22	0.30	0.22	3.83	1.81	0.67	0.05	0.00	0.00
BIC (33) (n, df_T)	0.23	0.32	0.25	3.78	2.10	1.06	0.05	0.00	0.00
BIC (33) (n, df_s)	0.22	0.32	0.23	3.83	1.97	0.79	0.05	0.00	0.00
BIC (34) (N, df_T)	0.26	0.35	0.29	4.04	2.32	1.24	0.06	0.00	0.00
BIC (34) (N, df_s)	0.23	0.34	0.23	4.04	2.31	0.87	0.08	0.00	0.00
BIC (33) (N, df_T)	0.26	0.35	0.29	4.04	2.32	1.24	0.06	0.00	0.00
BIC (33) (N, df_s)	0.23	0.34	0.23	4.04	2.31	0.87	0.08	0.00	0.00
SCAD with λ chosen by...									
Naïve GCV (21)	0.21	0.23	0.18	4.41	3.67	3.06	0.16	0.01	0.01
Fu GCV (50)	0.24	0.24	0.21	4.87	5.65	5.76	0.26	0.22	0.14
BIC (34) (n, df_T)	0.19	0.22	0.18	5.16	4.65	4.36	0.24	0.03	0.02
BIC (34) (n, df_s)	0.23	0.23	0.19	5.62	5.26	5.36	0.50	0.14	0.08
BIC (33) (n, df_T)	0.19	0.25	0.20	5.16	5.47	5.48	0.24	0.17	0.07
BIC (33) (n, df_s)	0.23	0.25	0.19	5.62	5.62	5.62	0.50	0.24	0.08
BIC (34) (N, df_T)	0.21	0.23	0.19	5.55	5.20	5.06	0.42	0.07	0.03
BIC (34) (N, df_s)	0.23	0.23	0.19	5.75	5.61	5.64	0.57	0.21	0.09
BIC (33) (N, df_T)	0.21	0.23	0.19	5.55	5.20	5.06	0.42	0.07	0.03
BIC (33) (N, df_s)	0.23	0.23	0.19	5.75	5.61	5.64	0.57	0.21	0.09
Oracle	0.13	0.19	0.15	6.00	6.00	6.00	0.00	0.00	0.00

Table 7: **Selected Models for Various Selection and Tuning Criteria for GEE with Exchangeable (Compound Symmetric) True Correlation.** The columns labeled C, O, U, and M show the proportion of trials in which the correct model, an overfit model, an underfit model, or a misfit model was selected by each method; see page 35 for definitions of these terms.

	Working Indep				Working AR-1				Working CS			
	C	O	U	M	C	O	U	M	C	O	U	M
Full Model	.00	1	.00	.00	.00	1	.00	.00	.00	1	.00	.00
Best Subset according to...												
Naïve AIC (33)	.40	.25	.16	.19	.37	.26	.16	.22	.34	.28	.24	.15
Naïve C_p (37)	.40	.25	.16	.20	.37	.25	.16	.23	.34	.28	.24	.15
Pan AIC (28)	.35	.31	.14	.21	.24	.52	.06	.19	.21	.56	.07	.17
Cantoni C_p (27)	.40	.25	.16	.20	.26	.47	.09	.19	.22	.53	.08	.18
BIC (33) (n)	.45	.13	.29	.15	.42	.06	.38	.15	.47	.06	.39	.09
BIC (33) (N)	.38	.04	.50	.09	.33	.00	.61	.07	.36	.01	.61	.03
BIC (34) (n)	.45	.13	.29	.15	.62	.12	.17	.10	.49	.13	.28	.11
BIC (34) (N)	.38	.04	.50	.09	.59	.02	.33	.07	.44	.03	.48	.06
LASSO with λ chosen by...												
Naïve GCV (21)	.01	.96	.00	.04	.00	1	.00	.00	.00	1	.00	.00
Fu GCV (50)	.03	.94	.01	.04	.01	.99	.00	.00	.00	1	.00	.00
BIC (34) (n, df_T)	.02	.94	.00	.05	.00	1	.00	.00	.00	1	.00	.00
BIC (34) (n, df_s)	.05	.90	.01	.04	.01	.99	.00	.00	.00	1	.00	.00
BIC (33) (n, df_T)	.02	.94	.00	.05	.00	1	.00	.00	.00	1	.00	.00
BIC (33) (n, df_s)	.05	.90	.01	.04	.01	.99	.00	.00	.00	1	.00	.00
BIC (34) (N, df_T)	.05	.90	.01	.06	.00	1	.00	.00	.00	1	.00	.00
BIC (34) (N, df_s)	.05	.88	.01	.07	.01	.99	.00	.00	.00	1	.00	.00
BIC (33) (N, df_T)	.05	.90	.01	.06	.00	1	.00	.00	.00	1	.00	.00
BIC (33) (N, df_s)	.05	.88	.01	.07	.01	.99	.00	.00	.00	1	.00	.00
SCAD with λ chosen by...												
Naïve GCV (21)	.18	.67	.03	.14	.05	.94	.00	.01	.02	.98	.01	.01
Fu GCV (50)	.27	.48	.15	.11	.50	.29	.20	.02	.67	.20	.13	.01
BIC (34) (n, df_T)	.34	.43	.11	.13	.19	.78	.01	.03	.15	.84	.02	.01
BIC (34) (n, df_s)	.32	.20	.41	.08	.46	.41	.11	.03	.55	.38	.07	.01
BIC (33) (n, df_T)	.34	.43	.11	.13	.49	.35	.12	.05	.56	.38	.06	.01
BIC (33) (n, df_s)	.32	.20	.41	.08	.50	.27	.20	.03	.65	.27	.08	.01
BIC (34) (N, df_T)	.34	.25	.31	.12	.38	.55	.03	.05	.35	.62	.02	.01
BIC (34) (N, df_s)	.31	.15	.49	.07	.49	.31	.18	.03	.64	.28	.09	.01
BIC (33) (N, df_T)	.34	.25	.31	.12	.38	.55	.03	.05	.35	.62	.02	.01
BIC (33) (N, df_s)	.31	.15	.49	.07	.49	.31	.18	.03	.64	.28	.09	.01
Oracle	1	.00	.00	.00	1	.00	.00	.00	1	.00	.00	.00

results, they are not discussed further. As mentioned earlier, there are many possible BIC-like tuning parameters for LASSO and SCAD, but the tables show only the performance of those the naïve, unweighted form (33). Performance of the corresponding criteria using (34) was also investigated but was very similar and is omitted to save space.

3.5.3 Binary-Response Simulation

Lastly, simulations were also performed for binary outcomes. This required some adjustments to create a comparable simulation setting. The β_j had to be smaller than before to avoid numerical instability, and the number of subjects and observations had to be considerably larger because of the smaller values of β_j and the smaller amount of information available per observation in a binary rather than Gaussian setting. For convenience, consider only the case in which the true correlation is exchangeable (compound symmetric); as before, working correlations include independent, exchangeable, and AR(1).

Table 8: **Performance Measures for Various Selection and Tuning Criteria for GEE with AR-1 True Correlation and Unbalanced Cluster Sizes.** Mean relative model error is defined in (61, and other measures are as in Table 2).

	Mean Model Error			Mean **Relative** Model Error			Proportion Correct Deletions			Proportion Wrong Deletions			Proportion Correct Models		
	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	0.29	0.16	0.24	1.00	0.74	1.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Best Subset according to...															
AIC (33)	0.25	0.15	0.21	0.81	0.67	0.97	5.12	4.96	5.03	0.22	0.16	0.19	0.32	0.28	0.33
C_p (naive)	0.25	0.15	0.21	0.81	0.67	0.97	5.12	4.95	5.01	0.22	0.15	0.19	0.32	0.28	0.32
AIC (Pan)	0.25	0.15	0.22	0.83	0.66	1.00	4.96	3.51	4.43	0.23	0.05	0.15	0.26	0.06	0.20
C_p (Cantoni)	0.25	0.19	0.21	0.81	0.87	0.98	5.12	5.75	5.36	0.22	0.53	0.26	0.32	0.40	0.39
BIC (light, (33))	0.23	0.16	0.21	0.77	0.74	0.95	5.68	5.84	5.75	0.38	0.35	0.34	0.46	0.57	0.56
BIC (heavy, (33))	0.23	0.19	0.22	0.82	0.88	1.01	5.90	5.96	5.92	0.53	0.58	0.49	0.44	0.41	0.49
BIC (light, (34))	0.23	0.20	0.21	0.77	1.03	0.95	5.68	5.79	5.71	0.38	0.43	0.31	0.46	0.51	0.54
BIC (heavy, (34))	0.23	0.21	0.22	0.82	1.11	1.00	5.90	5.91	5.92	0.53	0.54	0.47	0.44	0.47	0.51
LASSO with λ chosen by...															
Naïve GCV, df_S	0.23	0.16	0.21	0.78	0.78	0.96	3.42	4.30	3.63	0.06	0.00	0.01	0.01	0.15	0.05
Fu GCV	0.23	0.23	0.26	0.78	1.05	1.21	3.38	4.93	4.45	0.07	0.02	0.06	0.06	0.31	0.19
BIC (33) (n, df_T)	0.24	0.16	0.23	0.84	0.76	1.04	3.96	4.45	4.07	0.07	0.00	0.01	0.07	0.18	0.10
BIC (33) (n, df_s)	0.24	0.15	0.21	0.84	0.71	0.99	4.21	4.79	4.33	0.10	0.01	0.03	0.14	0.28	0.16
BIC (33) (N, df_T)	0.26	0.18	0.25	0.94	0.85	1.15	4.17	4.61	4.26	0.09	0.00	0.02	0.13	0.21	0.13
BIC (33) (N, df_s)	0.25	0.16	0.23	0.92	0.73	1.06	4.39	4.93	4.52	0.14	0.01	0.04	0.16	0.32	0.20
SCAD with λ chosen by...															
Naïve GCV, df_S	0.23	0.17	0.20	0.77	0.81	0.91	4.43	5.12	4.77	0.14	0.03	0.10	0.15	0.45	0.29
Fu GCV	0.24	0.22	0.24	0.79	1.05	1.10	5.00	5.80	5.86	0.23	0.14	0.46	0.29	0.69	0.46
BIC (33) (n, df_T)	0.22	0.16	0.20	0.73	0.76	0.93	5.27	5.52	5.34	0.23	0.04	0.17	0.38	0.58	0.41
BIC (33) (n, df_s)	0.23	0.15	0.21	0.79	0.68	0.97	5.67	5.69	5.62	0.40	0.09	0.33	0.40	0.66	0.45
BIC (33) (N, df_T)	0.23	0.19	0.22	0.79	0.87	1.04	5.59	5.68	5.63	0.32	0.06	0.28	0.42	0.66	0.47
BIC (33) (N, df_s)	0.24	0.16	0.22	0.85	0.71	1.02	5.80	5.81	5.82	0.53	0.12	0.41	0.35	0.71	0.47
Oracle (true subset)	0.16	0.11	0.15	0.53	0.51	0.71	6.00	6.00	6.00	0.00	0.00	0.00	1.00	1.00	1.00

Table 9: **Performance Measures for Various Selection and Tuning Criteria for GEE with Exchangeable True Correlation and Unbalanced Cluster Sizes.** Mean relative model error is defined in (61, and other measures are as in Table 8.

	Mean Model Error			Mean **Relative** Model Error			Proportion Correct Deletions			Proportion Wrong Deletions			Proportion Correct Models		
	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full Model	0.38	0.24	0.20	1.00	0.87	0.68	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Best Subset according to...															
AIC (33)	0.34	0.21	0.19	0.84	0.75	0.65	4.96	5.12	5.02	0.16	0.16	0.20	0.31	0.32	0.33
C_p (naive)	0.34	0.21	0.19	0.84	0.75	0.65	4.95	5.11	5.01	0.16	0.15	0.20	0.31	0.32	0.33
AIC (Pan)	0.35	0.22	0.19	0.88	0.78	0.64	4.72	3.72	3.42	0.18	0.07	0.09	0.21	0.07	0.03
C_p (Cantoni)	0.34	0.25	0.22	0.84	0.86	0.78	4.95	5.84	5.84	0.16	0.54	0.51	0.31	0.38	0.43
BIC (light, (33))	0.32	0.23	0.20	0.80	0.82	0.70	5.66	5.87	5.88	0.34	0.38	0.35	0.48	0.57	0.61
BIC (heavy, (33))	0.33	0.25	0.21	0.84	0.87	0.76	5.87	5.97	5.95	0.55	0.56	0.47	0.42	0.42	0.51
LASSO with λ chosen by...															
Naïve GCV, df_S	0.31	0.24	0.20	0.79	0.87	0.70	3.23	4.21	4.32	0.01	0.01	0.01	0.01	0.12	0.15
Fu GCV	0.32	0.32	0.26	0.80	1.15	0.93	3.31	4.66	4.87	0.01	0.05	0.01	0.06	0.20	0.31
BIC (33) (n, df_T)	0.32	0.25	0.21	0.83	0.87	0.73	3.99	4.31	4.47	0.01	0.01	0.01	0.07	0.13	0.17
BIC (33) (n, df_s)	0.32	0.23	0.19	0.81	0.82	0.67	4.22	4.54	4.71	0.03	0.03	0.01	0.14	0.20	0.27
BIC (33) (N, df_T)	0.34	0.27	0.23	0.90	0.95	0.80	4.20	4.41	4.58	0.02	0.01	0.01	0.11	0.15	0.21
BIC (33) (N, df_s)	0.32	0.24	0.21	0.84	0.86	0.72	4.42	4.68	4.87	0.06	0.04	0.03	0.16	0.23	0.29
SCAD with λ chosen by...															
Naïve GCV, df_S	0.33	0.23	0.20	0.81	0.83	0.69	4.25	5.02	5.17	0.09	0.08	0.04	0.12	0.35	0.46
Fu GCV	0.34	0.29	0.25	0.85	1.04	0.92	4.80	5.64	5.88	0.20	0.21	0.15	0.27	0.56	0.75
BIC (33) (n, df_T)	0.30	0.23	0.20	0.77	0.81	0.70	5.16	5.30	5.51	0.18	0.10	0.06	0.39	0.44	0.60
BIC (33) (n, df_s)	0.32	0.22	0.18	0.82	0.79	0.65	5.58	5.51	5.73	0.39	0.13	0.09	0.39	0.56	0.70
BIC (33) (N, df_T)	0.31	0.25	0.22	0.81	0.88	0.77	5.58	5.48	5.69	0.29	0.12	0.06	0.47	0.53	0.71
BIC (33) (N, df_s)	0.33	0.23	0.19	0.85	0.83	0.69	5.80	5.65	5.82	0.54	0.17	0.12	0.35	0.60	0.73
Oracle (true subset)	0.25	0.18	0.15	0.58	0.64	0.51	6.00	6.00	6.00	0.00	0.00	0.00	1.00	1.00	1.00

For each of 100 simulations of $n = 150$ subjects each with $J = 10$ observations, exchangeably correlated ($\rho \approx .3$) binary responses were generated in R using the `bindata` package of Leisch and Weingessel (2005; see Leisch, Weingessel, and Hornik, 1998, for background), by generating exchangeably correlated ($\rho = .7$) multivariate normal deviates and dichotomizing them at appropriate normal quantiles to create the desired marginal mean structure. This mean structure was $\mu_{ij} = P(Y_{ij} = 1) = g(-2 + \mathbf{X}_{ij}\beta)$ where $g(\cdot)$ is the logit link function, the x_{ij} are centered and scaled exchangeably correlated predictors as before, and β now is $[1, \frac{1}{3}, 0, 0, \frac{2}{3}, 0, 0, 0, \frac{1}{6}, 0]^T$.

As before, we considered the correct and incorrect deletion rates of the various selection criteria, as well as their success rate in finding the correct model. However, since model error as defined earlier no longer has a straightforward relationship to prediction error, we replace it with a simpler $MSE = \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2$. Table 10 shows results only for the naïve (unweighted) versions of GCV and BIC, because attempts to incorporate covariance into the binomial deviance function as in (48) led to very poor performance. Here, df_T denotes $\left(\left(\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i + n\boldsymbol{\Delta}\right)^{-1} \sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i\right)$ where $\mathbf{D}_i = \mathbf{A}_i \mathbf{X}_i$, $\mathbf{V}_i = \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}$, and $[\mathbf{A}_i]_{jj} = \hat{\mu}_{ij}(1 - \hat{\mu}_{ij})$, in the spirit of (22).

Table 10: **Performance Measures for Various Selection and Tuning Criteria for GEE with Binary Data and Exchangeable True Correlation.** Notation is defined on pages ?? and ??.

	Mean Model Error			Correct Deletions			Wrong Deletions			Correct Models		
	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS	Ind	Ar	CS
Full model	0.97	0.74	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Best AIC	0.78	0.56	0.57	5.04	5.28	5.30	0.28	0.27	0.32	0.32	0.37	0.42
Best AIC (Pan)	0.79	0.58	0.57	4.94	5.02	4.98	0.28	0.24	0.25	0.29	0.26	0.30
Best BIC (light)	0.68	0.55	0.60	5.81	5.76	5.80	0.52	0.47	0.53	0.49	0.51	0.50
Best BIC (heavy)	0.73	0.64	0.67	5.88	5.86	5.89	0.69	0.71	0.69	0.35	0.34	0.37
L_1, GCV, df_H	0.67	0.54	0.54	2.89	3.27	3.72	0.08	0.09	0.03	0.02	0.00	0.09
L_1, Light BIC (n), df_s	0.74	0.59	0.58	4.06	4.36	4.51	0.19	0.13	0.09	0.11	0.12	0.19
L_1, Light BIC (n), df_H	0.73	0.60	0.62	4.10	4.15	4.29	0.19	0.12	0.08	0.11	0.06	0.14
L_1, Heavy BIC (N), df_s	0.74	0.64	0.61	4.36	4.65	4.67	0.22	0.21	0.15	0.13	0.15	0.21
L_1, Heavy BIC (N), df_H	0.75	0.70	0.73	4.31	4.49	4.58	0.20	0.19	0.15	0.13	0.10	0.18
SCAD, GCV, df_H	0.81	0.58	0.54	4.89	5.29	5.38	0.31	0.29	0.29	0.27	0.40	0.46
SCAD, Light BIC (n), df_s	0.69	0.52	0.59	5.75	5.80	5.83	0.54	0.46	0.54	0.44	0.51	0.50
SCAD, Light BIC (n), df_H	0.69	0.52	0.59	5.75	5.78	5.84	0.54	0.45	0.54	0.44	0.52	0.50
SCAD, Heavy BIC (N), df_s	0.71	0.60	0.62	5.89	5.90	5.92	0.71	0.71	0.65	0.34	0.35	0.41
SCAD, Heavy BIC (N), df_H	0.71	0.60	0.62	5.89	5.89	5.91	0.71	0.70	0.66	0.34	0.35	0.40
Oracle (true model)	0.44	0.34	0.35	6.00	6.00	6.00	0.00	0.00	0.00	1.00	1.00	1.00

The results of the binary simulations were substantively similar to those for the normal case. As before, the best methods overall were direct use of light ($\log(n)$) BIC, SCAD with light BIC tuning, and naïve marginal testing. L_1 penalties usually selected overly large models but had the advantage of seldom committing wrong deletions. Regardless of method, the true model was not selected most of the time. As in Dziak and Li (2006), heavy ($\log(N)$) BIC seems to be too strong for the subtle effect sizes used in this simulation.

3.5.4 Summary

Various possibilities for GEE model selection, including adaptations of AIC, C_p, BIC, LASSO, and SCAD, all seem to work reasonably well. Various possibilities exist for incorporating correlation into the selection criterion, but fortunately the details of how this is done are not crucial for performance. SCAD with a light ($log(n)$) BIC worked quite well.

3.6 Proofs of Theorems 3.1 and 3.2

The asymptotic results in this and the following two chapters are very similar in structure, differing only in some technical details; proofs follow those of Fan and Li (2001). Proofs are given separately in each chapter to allow each chapter to be relatively self-contained in reasoning.

The arguments in this and Chapter **??** assume that q, d, and the unknown true $\boldsymbol{\beta}$ are fixed. This fixed-model assumption is somewhat controversial. It arguably makes selection too easy (see Fan and Peng, 2004; Leeb and Pötscher, 2005), since it implies that the active and inactive coefficients become indefinitely easier to tell apart, as the standard errors approach zero while the coefficients remain fixed. However, it is very common in the literature and is a good starting point, and will be loosened slightly in Chapter **??**.

Throughout this thesis I assume that the parameter space for $\boldsymbol{\beta}$ is compact. This may be somewhat unrealistic, but it is convenient so that we do not have to worry about whether the convergence of various functions is uniform (see Creel, 2006).

The proofs require the following regularity conditions:

RC 3.1 $\mathbf{s}_n(\boldsymbol{\beta})$ *and* $\mathbf{K}_n(\boldsymbol{\beta})$ *have continuous third derivatives in* β.

RC 3.2 $\mathbf{K}_n(\boldsymbol{\beta})$ *is positive definite with probability approaching one. There exists a non-random function* $\mathbf{K}_{n0}(\boldsymbol{\beta})$ *such that* $\|\mathbf{K}_n(\boldsymbol{\beta}) - \mathbf{K}_{n0}(\boldsymbol{\beta})\| \xrightarrow{p} 0$ *uniformly, and* $\mathbf{K}_{n0}(\boldsymbol{\beta}) > 0$ *for all* $\boldsymbol{\beta}$.

RC 3.3 *The* $\mathbf{s}_i(\beta) = \mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ *have finite covariance for all* $\boldsymbol{\beta}$.

Thus by RC 3.3 and the Central Limit Theorem, for any fixed $\boldsymbol{\beta}$, $\mathbf{s}_n(\boldsymbol{\beta})$ is $O_p(n^{1/2})$ and asymptotically normal.

RC 3.4 *The derivatives of* $\mathbf{K}_{n0}(\boldsymbol{\beta})$ *in* $\boldsymbol{\beta}$ *are* $O_p(1)$ $\forall \boldsymbol{\beta}$.

Suppose that $\mathcal{P}_n(\boldsymbol{\beta})$ in (54) has the form $\sum_{j=1}^n p(\boldsymbol{\beta}_{nj}, \lambda_{nj})$ where $p(\cdot)$ is nonnegative, symmetric at zero, and has a continuous second derivative except possibly at zero. Suppose also that the penalty $p(\cdot)$ and tuning parameters λ_{nj} (we allow λ to differ for each coefficient) are chosen such that

RC 3.5 $\sup_{j \in \mathcal{A}} |p'_{nj}(|\beta_{0j}|)| = O_p(n^{1/2} k_n^{-1})$

RC 3.6 $\sup_{j \in \mathcal{A}} \left(|p''_{nj}(\beta_{0j})| \right) = o_p(n k_n^{-1})$

RC 3.7 $p''_{nj}(\beta_j)$ *is a smooth function such that* $\exists c_1, c_2 > 0$ *such that*

$$\left| p''_{nj}(\theta_1) - p''_{nj}(\theta_2) \right| \leq c_2 |\theta_1 - \theta_2|$$

whenever $\theta_1, \theta_2 > c_1 \lambda_n$, *where* $\lambda_n \geq 0$ *is the tuning parameter.*

In (51), these hold for the L_1 penalty if the λ are $O_p(n^{-1/2})$ and for the SCAD penalty if the λ are $o_p(1)$.

Proof of Theorem 3.1

To begin, note that treating the $\mathbf{V}_i$ and $\mathbf{D}_i$ as fixed, we have (see p. **??**) $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{s}_n(\boldsymbol{\beta}) = \mathbf{K}_n$. Then by conditions 3.2 and 3.3 that for any fixed $\boldsymbol{\beta}$,

$$\begin{aligned} Q_n(\boldsymbol{\beta}) &\equiv \frac{1}{2n} \mathbf{s}_n^T(\boldsymbol{\beta}) \mathbf{K}_n^{-1}(\boldsymbol{\beta}) \mathbf{s}_n(\boldsymbol{\beta}) = \frac{1}{2n} \mathbf{s}_n^T(\boldsymbol{\beta}) \mathbf{K}_{n0}^{-1}(\boldsymbol{\beta}) \mathbf{s}_n(\boldsymbol{\beta}) + o_p(1), \qquad (62) \\ \dot{Q}_n(\boldsymbol{\beta}) &= \mathbf{s}_n(\boldsymbol{\beta}), \text{ and} \\ \ddot{Q}_n(\boldsymbol{\beta}) &= n \mathbf{K}_n(\boldsymbol{\beta}) = n \mathbf{K}_{n0}(\boldsymbol{\beta}) + o_p(n) \end{aligned}$$

It suffices to prove that $\forall \epsilon > 0$, with probability at least $1 - \varepsilon$, $\exists$ some large constant $\mathcal{C}_\varepsilon$ such that a local solution $\hat{\beta}_n$ exists in the interior of the ball

$$\left\{ \boldsymbol{\beta}_0 + n^{-1/2} \mathbf{u} : \ \|\mathbf{u}\| \leq C_\varepsilon \right\}$$

so that $\left\|\hat{\beta}_n - \boldsymbol{\beta}_0\right\|$ will be $O_p(n^{-1/2})$. (see Fan and Li, 2001) Let

$$D_n(\mathbf{u}) \equiv Q_n^*(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n^*(\boldsymbol{\beta}_0)$$

Then since the penalty is nonnegative,

$$\begin{aligned} D_n(\mathbf{u}) &= Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0) + k_n \sum_{j=1}^{d_n} \left\{ p_{nj}(\boldsymbol{\beta}_0 + n^{-1/2}u_j) - p_{nj}(\boldsymbol{\beta}_0) \right\} \\ &\leq Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - Q_n(\boldsymbol{\beta}_0) + k_n \sum_{j \in \mathcal{A}_n} \left\{ p_{nj}(\boldsymbol{\beta}_0 + n^{-1/2}u_j) - p_{nj}(\boldsymbol{\beta}_0) \right\} \end{aligned}$$

By condition 3.1, we can expand $Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u})$ as

$$\begin{aligned} Q_n(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) &= Q_n(\boldsymbol{\beta}_0) + n^{-1/2}\mathbf{u}^T \dot{\mathbf{Q}}_n(\boldsymbol{\beta}_0) + \frac{1}{2} n^{-1} \mathbf{u}^T \ddot{\mathbf{Q}}_n(\boldsymbol{\beta}_0)\mathbf{u} \\ &\quad + \frac{1}{6} n^{-3/2} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \mathbf{u}^T \ddot{\mathbf{Q}}_n(\boldsymbol{\beta}^*)\mathbf{u} \right\} \right)^T \mathbf{u} \\ &\equiv Q_n(\boldsymbol{\beta}_0) + (a) + (b) + (c) \end{aligned}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}$. Now

$$\|(a)\| = n^{-1/2}\mathbf{u}^T \mathbf{s}_n(\boldsymbol{\beta}_0) = O_p(1)\,\|\mathbf{u}\|$$

and

$$(b) = \frac{1}{2n}\mathbf{u}^T \ddot{\mathbf{Q}}_n(\boldsymbol{\beta}_0)\mathbf{u} = \mathbf{u}^T \mathbf{K}_{n0}(\boldsymbol{\beta}_0)\mathbf{u} + o_p(1)\,\|\mathbf{u}\|^2 .$$

Also, by condition 3.4, $\|(c)\| = O_p(n^{-1/2})\,\|\mathbf{u}\|^3$.

Next, use a second-order Taylor expansion of $p_{nj}(\beta_{0j} + n^{-1/2}u_j)$ around $p_{nj}(\beta_{0j})$, so that for $j \in \mathcal{A}$,

$$\begin{aligned} p_{nj}(\beta_{0j} + n^{-1/2}u_j) &= p_{nj}(\beta_{0j}) + n^{-1/2}u_j\; p'_{nj}(\beta_{0j}) + \frac{1}{2}\left(n^{-1}u_j^2\; p''_{nj}(\beta_j^*) \right) \\ &= p_{nj}(\beta_{0j}) + n^{-1/2}u_j\; p'_{nj}(\beta_{0j}) + \frac{1}{2}\left(n^{-1}u_j^2\; p''_{nj}\{\beta_{0j}\} \right)(1 + o(1)) \\ &= p_{nj}(\beta_{0j}) + n^{-1/2}u_j\; p'_{nj}(\beta_{0j}) + \frac{1}{2} n^{-1}u_j^2\; p''_{nj}(\beta_{0j}) + o(n^{-1}) \end{aligned}$$

where β_j^* is between β_{0j} and $\beta_{0j} + n^{-1/2}u_j$. That is,

$$k_n \sum_{j \in \mathcal{A}} p_{nj}(\boldsymbol{\beta}_0 + n^{-1/2}u_j) = k_n \sum_{j \in \mathcal{A}} p_{nj}(\boldsymbol{\beta}_0) + (d) + (e)$$

where $(d) = k_n n^{-1/2}\mathbf{u}^T \mathbf{p}_n^\dagger$, $(e) = \frac{1}{2}k_n n^{-1}\mathbf{u}^T \mathbf{p}_n^{\dagger\dagger}\mathbf{u}$, $\mathbf{p}_{nj}^\dagger = p''_{nj}(\beta_{0j})I(j \in \mathcal{A})$,

$\mathbf{p}_{nj}^{\dagger\dagger} = p''_{nj}(\beta_{0j})I(j \in \mathcal{A})$, and $I(\cdot)$ is the indicator function. By condition 3.5 then,

$$(d) = k_n n^{-1/2} O_p(k_n^{-1} n^{1/2}) \|\mathbf{u}\| = O_p(1) \|\mathbf{u}\|$$

and similarly $(e) = o_p(1) \|\mathbf{u}\|^2$ by condition 3.6.

Thus we have

$$\begin{aligned} D_n(\mathbf{u}) &= (a) + (b) + (c) + (d) + (e) \\ &= \mathbf{u}^T \mathbf{K}_{n0} \mathbf{u} + O_p(1) \|\mathbf{u}\| + o_p(1) \|\mathbf{u}\|^2 + o_p(1) \|\mathbf{u}\|^3 \end{aligned}$$

For sufficiently large n, $\mathbf{u}^T \mathbf{K}_{n0} \mathbf{u}$ dominates the other terms and is positive, by RC 3.2. Thus $D_n > 0$, so there is at least a local minimizer inside the ball.

Proof of Lemma 3.1

The proof of Lemma 3.1 follows directly from Theorem 3.1. Fix $\varepsilon = \frac{1}{2}\delta$. Then, since we assume that the true β_{0j} are fixed constants,

$$\begin{aligned} \Pr(\exists\, j \in \mathcal{A} : \hat{\beta}_j = 0) &\leq \Pr(\exists\, j \in \mathcal{A} : \left|\hat{\beta}_j - \beta_j\right| > \varepsilon) \\ &\leq \Pr(\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|^2 > \varepsilon^2) = o(1) \end{aligned}$$

Note that without the assumption of a fixed true β, this lemma is no longer trivial. This is related to the criticisms made by Leeb and Pötscher (2005) against consistent model selection methods. If we imagine that some of the nonzero $|\beta_{0j}|$ approach zero as n grows (think of a sequence of alternatives in a hypothesis testing problem), then depending on the rate at which this occurs we might not have sensitivity or consistency. In that case a non-consistent procedure like AIC (or perhaps LASSO or SCAD with GCV rather than a BIC-like tuning parameter (see Li et al., 2006)) might work as well or better in some sense than a consistent procedure. It is very difficult to say what asymptotic scenario is the fairest representation of a real data analyst's situation; much depends on the beliefs and goals of the analyst.

Proof of Lemma 3.2

By Theorem 3.1, it suffices to prove that for any sequence of local minimizers $\hat{\boldsymbol{\beta}}$ of Q_n^* such that $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\| = O_p(n^{-1/2})$, the inactive-coefficient estimates

$\hat{\boldsymbol{\beta}}_{\mathcal{N}}$ will be zero. Therefore, suppose

$$\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_0 = O_p(n^{-1/2}) \tag{63}$$

We can show that there then exists some small $k_0 > 0$ such that

$$Q_n^{\mathcal{P}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\mathcal{A}} \\ \mathbf{0}\end{bmatrix}\right) = \arg\min\left\{Q_n^{\mathcal{P}}\left(\begin{bmatrix}\hat{\boldsymbol{\beta}}_{\mathcal{A}} \\ \hat{\boldsymbol{\beta}}_{\mathcal{N}}\end{bmatrix}\right) : \left\|\hat{\boldsymbol{\beta}}_{\mathcal{N}}\right\| \leq k_0 n^{-1/2}\right\} \tag{64}$$

For (64) to hold, it is sufficient that for each $j \in \mathcal{N}$ we have $\operatorname{sgn}\left(\frac{\partial}{\partial \boldsymbol{\beta}_j} Q_n^{\mathcal{P}}(\boldsymbol{\beta}_j)\right) = -\operatorname{sgn}(\boldsymbol{\beta}_j)$ for $0 < \left|\boldsymbol{\beta}_j\right| < k_0 n^{-1/2}$, i.e., that the contribution to $Q_n^{\mathcal{P}}$ will be smaller for a $\beta_j = 0$ than for a β_j very near zero. (Intuitively, this is why we add a LASSO or SCAD penalty in the first place, and by assuming that expressions (55) and (56) hold, we are basically assuming that the penalty is strong enough.)

To check that (64) holds, do a second-order Taylor expansion of $\frac{\partial}{\partial \boldsymbol{\beta}} Q_n(\boldsymbol{\beta})$ about $\boldsymbol{\beta}_0$, so that

$$\begin{aligned}\frac{\partial Q^{\mathcal{P}}(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial Q(\boldsymbol{\beta}_0)}{\partial \beta_j} &+ \sum_{\ell=1}^{d} \frac{\partial^2 Q(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_\ell}\left(\beta_\ell - \beta_{0\ell}\right) \\ &+ \frac{1}{2}\sum_{\ell=1}^{d}\sum_{k=1}^{d}\left(\frac{\partial^3 Q(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_\ell \partial \beta_k}\right)\left(\beta_{n\ell} - \beta_{n0\ell}\right)\left(\beta_{nk} - \beta_{n0k}\right) \\ &+ k_n \frac{\partial p_{nj}(\beta_j)}{\partial \beta_j} \\ &\equiv (a) + (b) + (c) + (d)\end{aligned} \tag{65}$$

where $\boldsymbol{\beta}_0$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. By (62), (63), and Conditions 3.2, 3.3, and 3.4, $\|(a)\|$, $\|(b)\|$, and $\|(c)\|$ are $O_p(n^{1/2})$. Now divide both sides of (65) by $k_n \lambda_{nj}$ so

$$k_n^{-1}\lambda_{nj}^{-1}\frac{\partial Q^{\mathcal{P}}(\boldsymbol{\beta})}{\partial \beta_j} = O_p(k_n^{-1} n^{1/2} \lambda_{nj}^{-1}) + \lambda_{nj}^{-1}\frac{\partial p_{nj}(\beta_j)}{\partial \beta_j} \equiv (d) + (e)$$

By (56), (d) vanishes asymptotically; by (55), (e) does not vanish. Furthermore, $(d) = k_n \lambda_{nj}^{-1} p'_{nj}(\boldsymbol{\beta}_{nj}) \operatorname{sgn}(\boldsymbol{\beta}_j)$ so $\operatorname{sgn}((b)) = \operatorname{sgn}(\boldsymbol{\beta}_j)$ and so (64) holds.

Proof of Theorem 3.2

This proof follows that of Theorem 2 in Fan and Li (2001) (see also the asymptotic normality proofs in Liang and Zeger 1986 and Harris and Mátyás 1999 for unpenalized GEE and QIF). By Theorem 3.1 and Lemma 3.2, $wp \to 1$, there exists a sequence $\hat{\boldsymbol{\beta}}$ such that $\boldsymbol{\beta}_{n\mathcal{N}} = 0$ and $\partial Q_n^{\mathcal{P}}(\boldsymbol{\beta})/\partial\beta_j = 0$ for all $j \in \mathcal{A}$. Notice that

$$\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{A}}} Q_n(\boldsymbol{\beta}) = \mathbf{K}_{n\mathcal{A}}^T \mathbf{K}_n^{-1} \mathbf{s}_n = \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{s}_n + o_p(1)$$

letting $\mathbf{K}_{n\mathcal{A}} = \frac{1}{n}\partial \mathbf{s}_n/\partial \boldsymbol{\beta}_{\mathcal{A}} = \frac{1}{n}\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_{i\mathcal{A}}$ where the $\mathbf{D}_{i\mathcal{A}}$ are the appropriate submatrices of the derivative matrices $\mathbf{D}_i$, and assuming that $\mathbf{K}_{n\mathcal{A}} \xrightarrow{p} \mathbf{K}_{n0\mathcal{A}} > 0$ as in Condition 3.2. Similarly,

$$\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{A}} \boldsymbol{\beta}_{\mathcal{A}}^T} Q_n(\boldsymbol{\beta}) = n\mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} = n\mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} + o_p(n)$$

By Taylor expansions of $\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{A}}} Q_n(\boldsymbol{\beta})$ and $\frac{\partial}{\partial \boldsymbol{\beta}_{\mathcal{A}}} Q_n(\boldsymbol{\beta})$ about $\boldsymbol{\beta}_0$,

$$\begin{aligned}
\frac{\partial Q^{\mathcal{P}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial Q(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} + \sum_{\ell=1}^{d} \frac{\partial^2 Q(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&\quad + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \left\{ \sum_{\ell=1}^{d} \frac{\partial}{\partial \beta_\ell} \frac{\partial^2 Q(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&\quad + k_n \mathcal{P}_{n\mathcal{A}}^*(\boldsymbol{\beta}_0) + (k_n \mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0) + o_p(k_n)) \left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} \right) \\
&= \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{s}_n + n\mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
&\quad + k_n \mathcal{P}_{n\mathcal{A}}^*(\boldsymbol{\beta}_0) + k_n \mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0) \left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} \right) + O_p(1) + o_p(n^{-1/2} k_n)
\end{aligned} \tag{66}$$

using Conditions 3.1, 3.4 and 3.7, and Theorem 3.1. Here $\mathcal{P}^*$ and $\mathcal{P}^{**}$ are the derivatives of the penalty as defined on page **??**. Then

$$\begin{aligned}
-n^{-1/2} \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{s}_n = \sqrt{n} &\left(\mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} + \frac{k_n}{n} \mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0) \right) \left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}} \right) \\
&+ k_n n^{-1/2} \mathcal{P}_{n\mathcal{A}}^*(\boldsymbol{\beta}_0) + O_p(n^{-1/2}) + o_p(n^{-1} k_n)
\end{aligned}$$

By Condition 3.3 and the Central Limit Theorem, $\mathbf{C}_n^{-1} n^{-1/2} \mathbf{s}_n \xrightarrow{L} \mathrm{N}(\mathbf{0}, \mathbf{I}_s)$. By Conditions 3.2 and 3.6, $\left(\mathbf{K}_{n0\mathcal{A}} \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} + \frac{k_n}{n} \mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0) \right) = O_p(1)$.

Then using Slutsky's Theorem,

$$\begin{aligned}
&\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}\right) + \mathbf{b}_n = \\
&- \left(\mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}} + \frac{k_n}{n}\mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0)\right)^{-1} \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{C}_n \mathbf{C}_n^{-1} n^{-1/2} \mathbf{s}_n \\
&+ O_p(n^{-1/2}) + o_p(n^{-1}k_n) \\
&\xrightarrow{L} \mathrm{N}\left(\mathbf{0}, \left(\mathbf{H}_{n0\mathcal{A}} + \frac{k_n}{n}\mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0)\right)^{-1} (\mathbf{S}_{n0\mathcal{A}}) \left(\mathbf{H}_{n0\mathcal{A}} + \frac{k_n}{n}\mathcal{P}_{n\mathcal{A}}^{**}(\boldsymbol{\beta}_0)\right)^{-1}\right)
\end{aligned} \tag{67}$$

where $\mathbf{H}_{n0\mathcal{A}} = \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}}$ and $\mathbf{S}_{n0\mathcal{A}} = \mathbf{K}_{n0\mathcal{A}}^T \mathbf{K}_{n0}^{-1} \mathbf{C}_{n0} \mathbf{K}_{n0}^{-1} \mathbf{K}_{n0\mathcal{A}}$. Presumably, this variance would be smallest if we have modeled $\mathbf{R}$ well so that $\mathbf{K}_{n0}^{-1} \approx \mathbf{C}_{n0}$; otherwise we will have a higher variance than necessary, even in the limit.

References

A. Agresti. *Categorical data analysis*. Wiley Interscience, Hoboken, NJ, 2nd edition, 2002.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademai Kiado, Budapest, 1973.

A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96:939–955, 2001.

M. A. Babyak. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66:411–421, 2004.

G. A. Ballinger. Using generalized estimating equations for longitudinal data analysis. *Organizational research methods*, 7:127–150, 2004.

L. Breiman. Better subset regression using the nonnegative garotte. *Technometrics*, 37:373–384, 1995.

L. Breiman. Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383, 1996.

K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological methods and research*, 33: 261–304, 2004.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, New York, 2nd edition, 2002.

E. Cantoni, J. M. Flemming, and E. Ronchetti. Variable selection for marginal longitudinal generalized linear models. *Biometrics*, 61:507–514, 2005.

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

M. Creel. Econometrics, 2006. URL http://pareto.uab.es/mcreel/Econometrics/econometrics.pdf.

P. J. Diggle, P. Heagerty, K. Y. Liang, and S. L. Zeger. *Analysis of longitudinal data*. Oxford University Press, 2nd edition, 1998.

D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

D. D. Dunlop. Regression for longitudinal data: A bridge from least squares regression. *American Statistician*, 48(4):299–303, November 1994.

J. J. Dziak and R. Li. An overview on variable selection for longitudinal data. In D. Hong, editor, *Quantitative Medical Data Analysis*, chapter Submitted. World Sciences Publisher, Singapore, 2006.

J. J. Dziak, R. Li, and L. Collins. Review and comparison of some variable selection procedures for linear regression. Technical report, The Methodology Center, Penn State University, 2005.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

J. Fan. Comments on 'Wavelets in statistics: A review' by A. Antoniadis. *Journal of the Italian Statistical Association*, 6:131–138, 1997.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961, 2004.

J. J. Faraway. Practical regression and ANOVA using R, 2002. URL http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf.

D. P. Foster and E. I. George. The Risk Inflation Criterion for multiple regression. *Annals of Statistics*, 22:1947–1975, 1994.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–148, 1993.

W. J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

W. J. Fu. Penalized estimating equations. *Biometrics*, 59:126–132, 2003.

G. M. Furnival and R. W. Wilson. Regression by leaps and bounds. *Technometrics*, 16:499–512, 1974.

E. I. George. The variable selection problem. *Journal of the American Statistical Association*, 95:1304–1308, 2000.

M. J. Gurka. Selecting the best linear mixed model under REML. *American Statistician*, 60:19–26, 2006.

D. Harris and L. Mátyás. Introduction to the generalized method of moments estimation. In L. Mátyás, editor, *Generalized Method of Moments Estimation*, pages 3–29. Cambridge University Press, NY, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference and prediction*. Springer, New York, 2001.

R. M. Hauser. Better rules for better decisions. *Sociological Methodology*, 25: 175–183, 1995.

A. E. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

D. R. Hunter and R. Li. Variable selection using MM algorithms. *Annals of Statistics*, 33:1617–1642, 2005.

C. M. Hurvich and C. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76:297–307, 1989.

R. E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion. *Journal of the American Statistical Association*, 90:928–34, 1995.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378, 2000.

J. Kuha. AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research*, 33:188–229, 2004.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

F. Leisch and A. Weingessel. *bindata: Generation of Artificial Binary Data*, 2005. R package version 0.9-12.

F. Leisch, A. Weingessel, and K. Hornik. On the generation of correlated artificial binary data. Working paper series, sfb, adaptive information systems and modelling in economics and management science, Vienna University of Economics, http://www.wu-wien.ac.at/am, 1998.

C. Leng, Y. Lin, and G. Wahba. A note on the LASSO and related procedures in model selection. Technical Report 1091, Department of Statistics, University of Wisconsin, April 2004.

R. Li, J. Dziak, and Y. Ma. Nonconvex penalized least squares: Characterizations, algorithm and application. Manuscript., 2006.

K. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–676, 1973.

B. D. Marx. On ill-conditioned generalized estimating equations and toward unified biased estimation, 2000. URL citeseer.ist.psu.edu/289017.html.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, New York, 2nd edition, 1991.

A. J. Miller. *Subset selection in regression*. Chapman and Hall, New York, 2nd edition, 2002.

J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear*

statistical models. Richard D. Irwin, Inc, Homewood, Ill., 4 edition, 1996.

H. Ojelund, H. Madsen, and P. Thyregod. Calibration with absolute shrinkage. *Journal of Chemometrics*, 15:497–509, 2001.

W. Pan. Akaike's Information Criterion in generalized estimating equations. *Biometrics*, 57:120–125, 2001.

D. K. Pauler. The Schwarz criterion and related methods for normal linear models. *Biometrika*, 85:13–27, 1998.

A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

F. L. Ramsey and D. W. Schafer. *The Statistical Sleuth*. Duxbury, Pacific Grove, Ca., 2nd edition, 2002.

R. T. Rust, D. Simester, R. J. Brodie, and V. Nilikant. Model selection criteria: An investigation of relative accuracy, posterior probabilities, and combinations of criteria. *Management Science*, 41:322–333, 1995.

SAS Institute, Inc. *SAS/STAT (r) 9.1 User's Guide*. SAS Institute, Inc., Cary, NC, 2004.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978.

A. Sen and M. Srivastava. *Regression analysis*. Springer-Verlag, New York, 1990.

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

R. Shibata. Statistical aspects of model selection. In J. C. Willems, editor, *From Data to Model*, chapter 5. Springer-Verlag, New York, 1989.

B. Thompson. Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement*, 55:525–534, 1995.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

R. Tibshirani. A simple explanation of the Lasso and least angle regression (http://www-stat.stanford.edu/~tibs/lasso/simple.html), 2002.

H. Wang, R. Li, and C. Tsai. A consistent tuning parameter selector for SCAD. Manuscript, 2005.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

D. L. Weakliem. A critique of the Bayesian Information Criterion for model

selection. *Sociological Methods and Research*, 27:359–397, 1999.

R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–47, 1974.

Y. Yang. Prediction/estimation with simple linear models: Is it really that simple? Preprint, Institute for Mathematics and its Applications, University of Minnesota, 2004.

Y. Yang. Can the strengths of AIC and BIC be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005.

M. Yuan and Y. Lin. Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100: 1215–1225, 2005.

P. Zhao and B. Yu. On model selection consistency of Lasso. Technical report, University of California at Berkeley, November 2005.

B. Zheng. Summarizing the goodness of fit on generalized linear models for longitudinal data. *Statistics in Medicine*, 19:1265–1275, 2000.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the Lasso. Technical report, Stanford University, 2004.

W. Zucchini. An introduction to model selection. *Journal of Mathematical Psychology*, 44:41–61, 2000.