Biometrics 71, 731–740
September 2015
DOI: 10.1111/biom.12300

The Group Exponential Lasso for Bi-Level Variable Selection

Patrick Breheny*

Department of Biostatistics, University of Iowa, 145 N. Riverside Dr., N336 CPHB Iowa City, Iowa 52242, U.S.A.
**email:* patrick-breheny@uiowa.edu

Summary. In many applications, covariates possess a grouping structure that can be incorporated into the analysis to select important groups as well as important members of those groups. One important example arises in genetic association studies, where genes may have several variants capable of contributing to disease. An ideal penalized regression approach would select variables by balancing both the direct evidence of a feature's importance as well as the indirect evidence offered by the grouping structure. This work proposes a new approach we call the group exponential lasso (GEL) which features a decay parameter controlling the degree to which feature selection is coupled together within groups. We demonstrate that the GEL has a number of statistical and computational advantages over previously proposed group penalties such as the group lasso, group bridge, and composite MCP. Finally, we apply these methods to the problem of detecting rare variants in a genetic association study.

Key words: Group variable selection; Penalized regression; Rare variants.

1. Introduction

In regression problems, variables can often be thought of as grouped. This arises when an underlying factor contributes multiple individual predictors that are distinct, yet related. Common examples include a set of indicator variables for representing a single categorical variable, or a set of basis functions evaluated for a single continuous variable. Grouping can also be identified by scientific reasoning. Genetic variants may be thought of as grouped by the gene that they belong to; likewise, the expression of genes may be thought of as grouped by the pathways those genes belong to.

Penalized regression provides an attractive approach to variable selection, particularly in high-dimensional problems. Although the majority of the research in this area has dealt with individual (i.e., not grouped) variable selection, there has been a fair amount of recent work extending these approaches to grouped predictors. The most prominent method in this field is the group lasso (Yuan and Lin, 2006), which yields sparse solutions at the group level. The concept of bi-level selection—selecting not only the important groups, but important members within those groups—was introduced in Huang et al. (2009). An overview of penalized regression methods for both group selection and bi-level selection was provided in a recent review by Huang, Breheny, and Ma (2012).

The group bridge methodology proposed in Huang et al. (2009) suffers from a number of computational drawbacks that limit its applicability in practice, particularly for large data sets. Here, we propose a new method we call the group exponential lasso (GEL), in which the threshold for variable selection declines exponentially as evidence of its group's importance increases. We derive algorithms to efficiently fit these models, and make this methodology available via the `R` package `grpreg` (available at `cran.r-project.org/package=grpreg`). The GEL has a number of appealing properties in terms of estimation accuracy as well as individual- and group-level variable selection, and can be scaled up to deal with very large problems, which make it particularly well-suited for the problem of detecting rare variants in genetic association studies.

We define the GEL, illustrate some of its mathematical properties, and develop algorithms for fitting the proposed model in Section 2. We demonstrate that the GEL performs well in simulation, outperforming other group penalization methods in a number of scenarios, in Section 3. Finally, we apply the GEL to data from the 1000 Genomes Project and illustrate its advantages over competing methods in the analysis of genetic association studies involving rare variants.

2. Group Exponential Lasso

We consider models in which the relationship between the outcome and the explanatory variables is specified in terms of a linear predictor $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \beta_0 + \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j, \tag{1}$$

where $\mathbf{X}_j$ is the portion of the design matrix formed by the predictors in the jth group and the vector $\boldsymbol{\beta}_j$ consists of the associated regression coefficients. Letting K_j denote the number of members in group j, $\mathbf{X}_j$ is an $n \times K_j$ matrix with elements (x_{ijk}), the value of kth covariate in the jth group for the ith subject. Covariates that do not belong to any group may be thought of as a group containing a single member. The total number of explanatory variables is $p = \sum_j K_j$.

The problem of interest involves estimating a vector of coefficients $\boldsymbol{\beta}$ defined by minimizing an objective function $Q(\boldsymbol{\beta})$ composed of a loss function L that quantifies the discrepancy between y_i and η_i combined with a penalty P that encourages

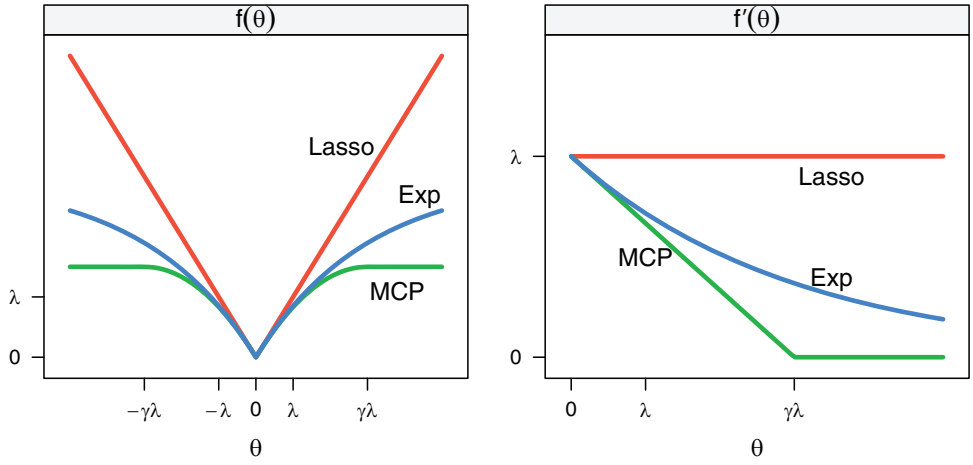

Figure 1. Lasso, MCP, and exponential penalty functions (left) and penalization rates (right). Note that none of the penalties are differentiable at $\beta_j = 0$, but that all three have finite directional derivatives everywhere.

sparsity and prevents overfitting:

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) + P(\boldsymbol{\beta}|\lambda), \tag{2}$$

where the regularization parameter λ controls the tradeoff between loss and penalty.

To ensure that the penalty is invariant to scale, covariates are standardized prior to fitting such that $\sum_i x_{ijk} = 0$ and $n^{-1}\sum_i x_{ijk}^2 = 1$. We assume without loss of generality that the covariates are standardized in this way during the model fitting process and then transformed back to the original scale once all models have been fit. This section focuses primarily on linear regression where $L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/2n$ and where we also assume, again without loss of generality, that the response has been centered such that $\sum_i y_i = 0$. In this case $\widehat{\beta}_0 = 0$ and may be ignored. Generalized linear models are considered in Section 2.4.

In the bi-level selection case, the penalty should incorporate the information contained in the grouping structure, which reflects a prior belief that important features are likely to be clustered to some extent in these groups. This involves combining penalties at the group and individual variable levels. A simple way of combining these penalties is by adding them, as in the following:

$$P(\boldsymbol{\beta}|\lambda) = \alpha\lambda \sum_j \sum_k \left|\beta_{jk}\right| + (1-\alpha)\lambda \sum_{j=1}^{J} \left\|\boldsymbol{\beta}_j\right\|. \tag{3}$$

This approach, in which the penalty consists of the lasso penalty plus the group lasso penalty, is known as the *sparse group lasso*, originally proposed in Wu and Lange (2008) and further developed in Simon et al. (2013). Here, the α parameter controls the tradeoff between the two penalties, with the lasso and group lasso representing special cases ($\alpha = 1$ and $\alpha = 0$, respectively).

Alternatively, one may combine group and individual variable penalties in a hierarchical framework (Breheny and Huang, 2009):

$$P(\boldsymbol{\beta}|\lambda) = \sum_{j=1}^{J} f_o\left\{\sum_{k=1}^{K_j} f_i(\beta_{jk})\right\}, \tag{4}$$

where f_o and f_i are penalties—the outer and inner penalties, respectively—both of which, in general, may involve λ as well as other tuning parameters. Although this framework is very general, it may be *too* general, in the sense that not all combinations of outer and inner penalties necessarily produce sensible models. A less general framework was proposed in Huang et al. (2012):

$$P(\boldsymbol{\beta}|\lambda) = \sum_{j=1}^{J} f\left(\left\|\boldsymbol{\beta}_j\right\|_1|\lambda\right). \tag{5}$$

If $f(\cdot)$ is a concave function on $[0, \infty)$, the model will produce solutions with grouping properties; the authors refer to this class of penalties as *concave 1-norm group penalties*.

Here, we propose a method for bi-level variable selection called the *group exponential lasso*, in which f is a concave exponential function. We define the exponential function and relate it to other commonly used penalty functions in Section 2.1, then define the GEL and discuss its properties in Section 2.2.

2.1. *The Exponential Penalty*

Consider the following function, defined on $[0, \infty)$, which we call the *exponential penalty*:

$$f(\theta|\lambda, \tau) = \frac{\lambda^2}{\tau}\left\{1 - \exp\left(-\frac{\tau\theta}{\lambda}\right)\right\}. \tag{6}$$

The logic behind the penalty can be seen in Figure 1, which contrasts the penalization rate of the exponential penalty with that of the lasso and the minimax concave penalty, or MCP (Zhang, 2010). MCP and the exponential penalty begin by applying the same rate of penalization as the lasso, but

continuously relax that penalization. MCP relaxes the rate of penalization linearly, and thus results in $f'(\theta) = 0$ for all $|\theta| > \gamma\lambda$, where γ is an additional tuning parameter of MCP playing a role similar to τ in (6). The exponential penalty, on the other hand, allows the penalty to decay exponentially, approaching $f'(\theta) = 0$ asymptotically but never reaching it. The diminishing rate of penalization is an attractive property; as discussed in Fan and Li (2001), it leads to the estimator $\widehat{\boldsymbol{\beta}}$ being nearly unbiased given a large enough sample size. The lasso does not have this property, and introduces significant bias toward zero for large regression coefficients.

Like the MC penalty and unlike the lasso, the exponential penalty is not convex. However, like the MC penalty, there are reasonable conditions under which the objective function is convex. These conditions are presented in the following proposition, which applies to the least squares loss function and the (ungrouped) exponential penalty, with objective function

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} f(\beta_j|\lambda, \tau), \tag{7}$$

where $f(\beta_j|\lambda, \tau)$ is given in (6).

PROPOSITION 1. *Let ξ_* denote the minimum eigenvalue of $n^{-1}\mathbf{X}'\mathbf{X}$. Then objective function* (7) *is strictly convex if $\tau < \xi_*$.*

The above proposition has the following corollary:

COROLLARY 1. *Let $Q(\beta_j)$ denote* (7) *considered as a function only of β_j, with all other coefficients fixed. Then $Q(\beta_j)$ is strictly convex if $\tau < 1$.*

In other words, provided that the rate of exponential decay, τ, is not too sharp, we will have a stable objective function with a unique global minimum. Furthermore, since the penalty function is separable and the objective function is convex in each coordinate dimension, we may apply a coordinate descent approach to solve for $\boldsymbol{\beta}$ and this approach is guaranteed to converge to the minimum. Of course, in high dimensions ($p \geq n$), the minimum eigenvalue ξ_* will be zero, so strict global convexity is not possible. However, local convexity (convexity over a subset of the coefficients) may still apply; in this case, ξ_* in Proposition 1 denotes the minimum eigenvalue of $n^{-1}\mathbf{X}_A'\mathbf{X}_A$, where A denotes the subset of active covariates. See Breheny and Huang (2011) for further discussion of local convexity.

The focus of this paper is on using the exponential penalty for grouped regularization, and thus a detailed study of the properties of the estimator for the ungrouped case is outside our scope. However, given the similarity of the penalty to the MC penalty, the estimators are likely to have similar estimation properties. The exponential penalty is also similar to the bridge penalty, in which the penalty applied to the jth covariate is $|\beta_j|^q$. Like the exponential penalty, the penalization rate of the bridge penalty with $q < 1$ diminishes gradually toward zero as β increases. The most important difference between the exponential penalty and the bridge penalty is that, as $\beta \to 0$, the penalization rate of the exponential penalty is stable and approaches to the rate applied by the lasso. For the bridge penalty, on the other hand, $\lim_{\beta\to 0^+} f'(\beta) = \infty$. This singularity at zero leads to considerable practical difficulties in working with bridge (and group bridge) penalties, as these penalties are not directionally differentiable at 0. The exponential penalty shares many of the same aspects as the bridge penalty, but offers considerable advantages from an algorithmic perspective, as it allows for convex objective functions and remains directionally differentiable everywhere.

2.2. *The Group Exponential Lasso*

Our main motivation for proposing the exponential penalty is for use as an outer penalty in the group regularization framework given by (4), or more specifically, as the concave penalty f in (5). Note that the partial derivative of the penalty function in (4) with respect to the jkth covariate, which we denote Δ_{jk}, is

$$\Delta_{jk} = f_o'\left\{\sum_{\ell=1}^{K_j} f_i(\beta_{j\ell})\right\} f_i'(\beta_{jk}). \tag{8}$$

In other words, the penalization rate applied to a coefficient depends on two factors. The first is the magnitude of the coefficient itself; this is the part given by $f_i'(\beta_{jk})$ and was plotted in Figure 1. The other part of the equation is governed by the magnitude of the other coefficients in its group. The main idea is this: suppose we are given two features which display equal association with the outcome. One of the features is located in a group in which several other features are important, the other resides in a group for which none of the other features are important. Given this information, it makes sense to select the first feature and decide that the second association is spurious. For example, if the features are variant alleles in a genetic association study, signal from a variant in a gene seemingly unrelated to the outcome is much more likely to be spurious than a variant in a gene housing other variants known to be associated with the disease.

The motivation behind the exponential penalty is to have the penalization applied to a coefficient β_{jk} decay exponentially as group j grows in importance, with the parameter τ controlling the rate of that decay, and thus, the strength of grouping. To see this, let $P(\boldsymbol{\beta})$ take the form (5), with f the exponential penalty given by (6); we refer to the resulting penalty as the *group exponential lasso*. In this case,

$$\Delta_j = \lambda \exp\left\{-\frac{\tau}{\lambda}\left\|\boldsymbol{\beta}_j\right\|_1\right\}; \tag{9}$$

note that we have dropped the subscript k from the expression, since the partial derivative is Δ_j for all members of the jth group. Again, with the ordinary lasso penalty this partial derivative is λ for all coefficients; with the GEL this penalization rate may be modified by the relevance of the group to the outcome. If no members of group j have been selected, $\left\|\boldsymbol{\beta}_j\right\|_1 = 0$ and all members of the group are penalized at the full rate λ. However, if some members are nonzero, then $\left\|\boldsymbol{\beta}_j\right\|_1 > 0$ and the penalization rate will be diminished. In a sense, the effect is similar to the adaptive lasso (Zou, 2006), although here the modifications to λ arise naturally from the

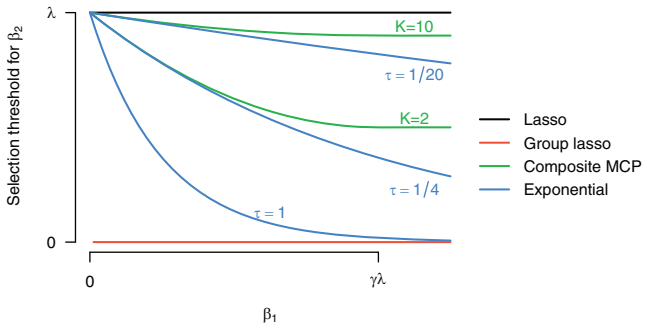

Figure 2. Effect of changing β_1 on the selection threshold for β_2 for two components in the same group.

hierarchical structure of the penalty rather than being imposed externally.

We refer to the phenomenon whereby the penalty applied to a coefficient is diminished if it is grouped with other important predictors as *coupling*. With the exponential penalty, the strength of coupling is determined by τ. Thus, in what follows we refer to τ as the *coupling parameter*.

To make the notion of coupling more concrete, suppose we have a group with two coefficients, β_1 and β_2 (we momentarily drop the group-level subscript here for the sake of simplicity). Now consider the effect that changing β_1 has on the selection threshold for β_2. Letting z denote the unpenalized solution for β_2 (i.e., the value that minimizes the loss function with the other β's fixed at a given value), we define the "selection threshold" as $\inf\{z : \widehat{\beta}_2 \neq 0\}$. In other words, the selection threshold is the minimum association that $\mathbf{x}_2$ must have with the outcome in order for it to be selected.

The choice of inner and outer penalty in (4) determines the extent of coupling; this is plotted in Figure 2. For example, suppose f_i and f_o are both lasso penalties. In this case, the penalty reduces to that of the conventional lasso, and no coupling takes place: the selection threshold for β_2 remains λ regardless of the value of β_1. At the other extreme, consider the group lasso of Yuan and Lin (2006). Here, the selection threshold for β_2 is λ when $\beta_1 = 0$, but if β_1 is nonzero, the selection threshold for β_2 instantly drops to 0. From a selection standpoint, this represents absolute coupling: it is not possible to select β_1 without β_2. For the sake of clarity, the sparse group lasso (SGL) is not represented on the plot, but its coupling profile is similar to the group lasso, albeit with the selection threshold dropping from λ to $\alpha\lambda$ instead of to zero when $\beta_1 \neq 0$. Thus, solutions are partially coupled in SGL as in the composite MCP and exponential approaches, although in a discontinuous, step-function fashion.

The composite MCP was proposed in Breheny and Huang (2009) as a compromise between these two extremes—to allow the selection of β_{jk} to be influenced by the other coefficients in group j, but in such a manner that β_{jk} must still show an independent association with the outcome to be included in the model. As the figure illustrates, this behaves reasonably for small group sizes. For example, in a two-component group ($K = 2$), the selection threshold for β_2 drops to $\lambda/2$ when β_1 is large ($\geq \gamma\lambda$). However, for the composite MCP, the extent of coupling is limited by the group size K: when β_1 is large, the selection threshold for β_2 drops to $\lambda(K-1)/K$; Figure 2 illustrates this for $K = 10$. In this case, when feature 2 is in a group with another feature strongly associated with the outcome, but the other eight are not associated with the outcome, its selection threshold is only 10% lower than that of a feature in a group lacking any associations with the outcome.

It is possible, of course, that this degree of coupling is appropriate for the problem at hand. However, for many applications it represents a rather insubstantial amount of coupling. For the genetic association application of Section 4, for example, we would expect the solution to be highly sparse across groups, as only a very small number of genes are likely to be related to the outcome. Furthermore, many of the groups are large, consisting of dozens or even hundreds of variants. For these groups, the composite MCP provides essentially no coupling—when a variant or two is chosen from these groups, the effect on selection for the other variants in the group is minimal. This is inconsistent with the nature of the problem. Variants tend to be functionally related, so that if one variant of a gene impacts an outcome, it is quite likely that others variants of the same gene also have an impact.

A limitation of the composite MCP is the fact that there is no parameter through which the analyst can control the strength of coupling to match the problem at hand. The GEL, on the other hand, does provide a coupling parameter in τ. Figure 2 illustrates the effect that varying this parameter has on the coupling between two elements in a group. As τ increases, so does the degree of coupling. With $\tau = 1/20$, the selection threshold for β_2 is only minimally affected by β_1; with $\tau = 1$, the selection threshold drops sharply as β_1 enters the model.

The notion of τ as a coupling parameter can also be justified in the sense that, as $\tau \to 0$, the GEL becomes equivalent to the standard lasso. This phenomenon is formally stated in Proposition 2.

PROPOSITION 2. *Letting f denote the exponential penalty defined in* (6),

$$\lim_{\tau \to 0} \sum_{j=1}^{J} f\left(\left\|\boldsymbol{\beta}_j\right\|_1\right) = \lambda \sum_{j=1}^{J} \sum_{k=1}^{K_j} \left|\beta_{jk}\right|$$

In principle, we could apply the idea of the group exponential penalty to (4) with any f_i as an inner penalty. For the sake of simplicity, we focus here on using the lasso as the inner penalty f_i, and refer to this approach as the GEL. We note, however, that the algorithms described in Section 2.3 apply other f_i such as MCP and SCAD; indeed, the extensions are rather straightforward and we briefly explored the properties of a "group exponential MCP," with MCP as the inner penalty. We found the empirical performance of this estimator to be relatively similar to that of GEL, although the estimator may be worth exploring further; we return to this point in Section 5.

2.3. *Model Fitting*

We apply the local coordinate descent approach proposed in Breheny and Huang (2009) to solve for the minimum of the

GEL objective function. Like other coordinate descent algorithms, we optimize the objective function with respect to a single parameter β_{jk} at a time. Coordinate descent algorithms are typically advantageous in settings where one-dimensional updates have simple, closed-form solutions. The GEL objective function lacks a closed-form one-dimensional solution; nevertheless, it may be locally approximated by a majorizing function that does have a closed form solution. he theory underlying majorization–minimization algorithms then ensures that the proposed algorithm possesses the descent property and is guaranteed to converge.

A detailed description of the algorithm is provided in the supplementary materials, but the main idea is to take a first-order Taylor series expansion of the penalty function about the current value of $\boldsymbol{\beta}$. This results in a linear approximation proportional to the lasso penalty, allowing efficient updating of model coordinates according to

$$\beta_{jk} \leftarrow S(z_{jk}|\tilde{\Delta}_j),$$

where S is the soft-thresholding operator (Donoho and Johnstone, 1994), z_{jk} is the unpenalized solution (i.e., the value of β_{jk} that minimizes $L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ with all other β values held constant), and $\tilde{\Delta}_j$ is given by (9) and evaluated at the current value of $\boldsymbol{\beta}_j$.

The local coordinate descent algorithm has two very attractive properties. First, because no cumbersome matrix operations are involved, iterations can be computed quite rapidly. Specifically, updating each parameter requires only $O(2n)$ operations, and thus each full iteration, cycling over the entire vector of parameters, can be accomplished in $O(2np)$ operations. Because the algorithm in linear in p, it can be scaled up to handle very high-dimensional problems. The other advantage of the local coordinate descent algorithm is its stability: Proposition 3 establishes that it is guaranteed to decrease the objective function with each iteration.

PROPOSITION 3. *Let $\{\boldsymbol{\beta}^{(k)}\}$ denote the sequence of coefficients produced at each iteration of the local coordinate descent algorithm for fitting group exponential lasso models. For all $k = 0, 1, 2, \ldots,$*

$$Q(\boldsymbol{\beta}^{(k+1)}) \leq Q(\boldsymbol{\beta}^{(k)}).$$

Furthermore, the sequence $\{\boldsymbol{\beta}^1, \boldsymbol{\beta}^2, \ldots\}$ is guaranteed to converge to a stationary point of $Q(\boldsymbol{\beta})$.

Proposition 3 works because the GEL penalty is concave on $[0, \infty)$, and therefore a linear (lasso) penalty provides a majorizing approximation that is both easy to solve and drives the objective function "downhill." It is worth noting, however, that since the objective function is not itself convex, the above proposition does not rule out convergence to local minima.

As has been described elsewhere (Friedman, Hastie, and Tibshirani, 2010; Breheny and Huang, 2011), we are usually interested in obtaining $\widehat{\boldsymbol{\beta}}$ for a path of λ values and then choosing among those models using either cross-validation or some form of information criterion. The continuity of these solution paths allows the algorithm to efficiently choose initial values that are never too far from the solution, a phenomenon known as "warm starts." In this regard, the GEL penalty has considerable computational advantages over the group bridge penalty (Huang et al., 2009), while retaining similar estimation properties. Both penalties are concave 1-norm group penalties of form (5), and as noted in Section 2.1, the exponential penalty is similar in shape and functional form to the bridge penalty. However, the group bridge penalty suffers from singularities at $\boldsymbol{\beta}_j = 0$, which cause a number of computational problems (detailed in Breheny and Huang, 2009) and prevent taking advantage of warm starts. The GEL penalty, on the other hand, presents none of these difficulties, and as we will see in Section 4, can be stably and efficiently scaled up to handle very large problems. In particular, the data sets in Section 4 contain roughly 25,000 variables and GEL models may be fit to the data in 2–4 seconds on a standard desktop computer. In comparison, the SGL requires several minutes to fit a model to the same data, and group bridge models are not computationally feasible to fit to data of this scale.

2.4. *Generalized Linear Models*

The penalties and algorithms we have described are readily extended to models with loss functions other than least squares. In particular, for generalized linear models (GLMs; McCullagh and Nelder, 1989) the loss function is the negative log likelihood from an exponential family. The typical approach to fitting such models is the iteratively reweighted least squares (IRLS) algorithm, which involves making a quadratic approximation to the loss function, solving the quadratic approximation, and repeating until convergence. It is straightforward to incorporate the local coordinate descent algorithm described above into the IRLS algorithm, though some subtle issues arise concerning the relationship between the coupling parameter τ and convexity.

In the IRLS algorithm, a working response (or adjusted response) $\tilde{\mathbf{y}}$ and diagonal matrix of weights $\mathbf{W}$ are calculated based on a Taylor series expansion about the current estimates, $\widetilde{\boldsymbol{\beta}}$, so that

$$L(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \approx \frac{1}{2n}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})$$

in the neighborhood of $\widetilde{\boldsymbol{\beta}}$. The linear approximation of the penalty in the local coordinate descent algorithm is unchanged; thus, letting z_{jk} represent the value of β_{jk} that minimizes this quadratic approximation, the updating step for a GLM becomes

$$\beta_{jk} \leftarrow \frac{S(z_{jk}|\tilde{\Delta}_j)}{v_{jk}},$$

where $v_{jk} = n^{-1}\mathbf{x}_{jk}'\mathbf{W}\mathbf{x}_{jk}$. It is worth noting that, analogous to Proposition 3, this approach is guaranteed to drive the quadratic approximation downhill, but like the IRLS algorithm itself, is not guaranteed to converge monotonically with respect to the actual objective function.

A subtle issue that presents itself for the GEL (indeed, for any nonconvex penalty) is the fact that the curvature of the loss function is distorted by $\mathbf{W}$. Thus, the conclusion in Corollary 1 no longer holds, as the minimum eigenvalue in Proposition 1 now depends on $\mathbf{W}$, which is not constant.

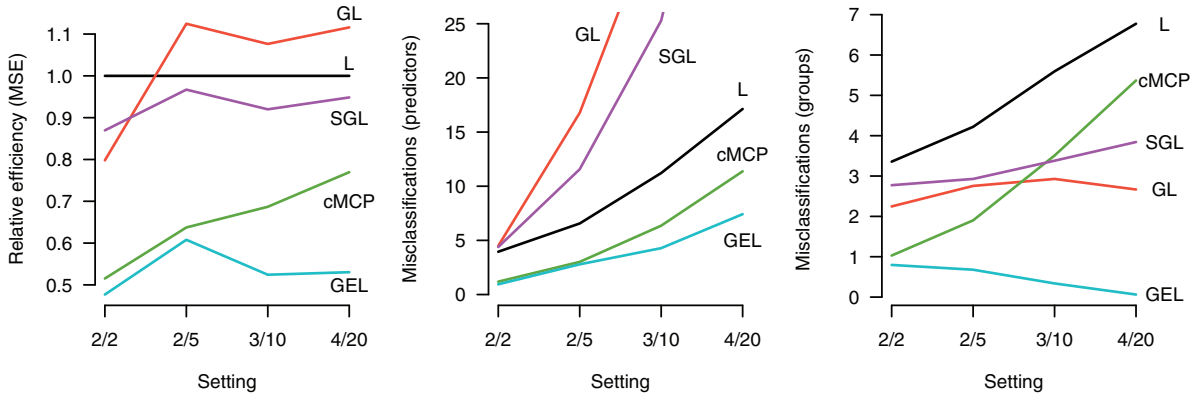

Figure 3. Effect of group size on performance. The "Setting" is given as {Number of nonzero coefficients} / {Number of variables per group}. All settings contain $J = 10$ groups and a sample size of $n = 100$; results are averaged over 600 independently generated data sets. For the leftmost plot, estimation efficiency is given as mean squared error (MSE) relative to the lasso. For the center and right plots, "misclassification" means that a predictor/group was selected despite having a true coefficient of zero, or vice versa. Methods: Lasso (L), Group lasso (GL), Sparse group lasso (SGL), composite MCP (cMCP), group exponential lasso (GEL).

This is not a fatal problem—there still exist values of τ that produce convex objective functions—but does represent something of an inconvenience to a user of these methods. In linear regression, the coupling parameter τ has a rather attractive, interpretable scale: $\tau = 0$ represents no coupling, equivalent to the traditional lasso, while $\tau = 1$ represents the maximum allowable degree of coupling before the objective function splinters into an unmanageable mess of local minima.

We would like to retain this convenient $\tau \in (0, 1)$ scale so that the analyst does not have to recalibrate the meaning of τ depending on the type of model that he or she is fitting. This can be accomplished in a straightforward manner by applying the idea of adaptive rescaling introduced in Breheny and Huang (2011); see supplementary materials for details. The primary purpose of this is to allow τ to remain unaffected by the weights and thereby take on a consistent interpretation in terms of coupling and convexity across different classes of models.

3. Simulation Studies

In this section, we evaluate the performance of the GEL using simulated data. We compare GEL against various other penalized regression methods, including the lasso (Tibshirani, 1996), MCP (Zhang, 2010), group lasso (Yuan and Lin, 2006), and composite MCP (Breheny and Huang, 2009). Lasso and MCP are completely uncoupled methods, while group lasso is completely coupled. Composite MCP is a bi-level selection method, like GEL, designed to strike a compromise between those two extremes.

A sample size of $n = 100$ was used throughout, while the number of features and groups is varied. In all of the studies, an external validation set, also of size $n = 100$, was used to choose the regularization parameter λ for each method. For MCP and composite MCP, γ was set to 3. Throughout, we fix the signal-to-noise ratio of the generating model at 1 (i.e., the R^2 of the true model is 0.5). Covariate values and error terms were generated independently from the standard normal distribution. Equivalent simulations were carried out for logistic regression; the results were similar and are included in the supplemental materials.

3.1. *Varying Group Size*

In this section, we aim to compare the various group penalization methods. Here, all of the nonzero coefficients reside within a single group. Thus, the grouping information is useful and group penalization methods should outperform the lasso. Here we fix J, the number of groups, at 10 and vary the number of elements within each group among $\{2, 5, 10, 20\}$ (i.e., the total number of predictors varies from 20 to 200).

The number of nonzero coefficients in the non-null group varies correspondingly among $\{2, 2, 3, 4\}$. As discussed in Section 2.2, the coupling strength of the composite MCP diminishes as group size increases; we anticipate that GEL will outperform composite MCP at larger group sizes. For the GEL, we set τ equal to 1/3 (the effect of varying τ will be explored in Section 3.2).

The results of this simulation are presented in Figure 3, which evaluates performance in terms of estimation efficiency and selection accuracy at both the individual predictor and group levels. Let us first considering estimation efficiency. As one would expect, the group penalization methods generally outperform the lasso. The exception to this statement is that the group lasso does not outperform the lasso in the presence of incomplete grouping (i.e., where groups have both zero and nonzero coefficients). When grouping is incomplete, the bi-level selection methods produce more accurate estimates than both lasso and group lasso. The hierarchical approaches, in turn, are substantially more accurate than the SGL. Composite MCP and GEL are similar for small groups (five or fewer predictors per group), but GEL becomes substantially more efficient when group sizes are large. In particular, GEL achieves an estimation efficiency 32% better than that of composite MCP in the "20 predictors per group" setting.

Similar remarks hold for variable selection accuracy. The group lasso performs quite poorly in the incompletely grouped

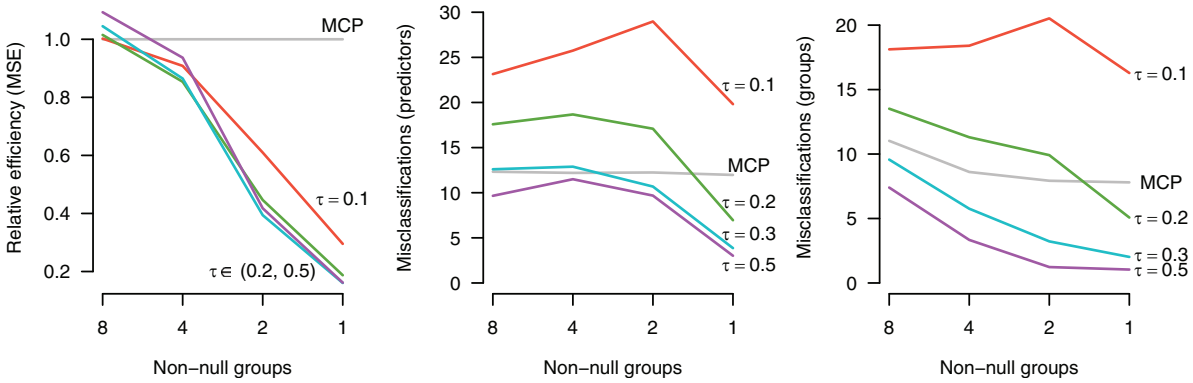

Figure 4. Effect of degree of grouping on performance. In each setting, there are eight non-null predictors, distributed among a variable number of groups ranging from 8 to 1. All settings contain $J = 100$ groups of size 10 and a sample size of $n = 100$; results are averaged over 600 independently generated data sets. For the leftmost plot, estimation efficiency is given as mean squared error (MSE) relative to MCP. For the center and right plots, "misclassification" means that a predictor/group was selected despite having a true coefficient of zero, or vice versa.

settings, as it is unable to select variables within groups. Although the SGL is able to select variables within groups, its selection accuracy is not much better than group lasso. Once again, composite MCP and GEL are similar for small groups, but with 20 predictors per group, composite MCP misclassified 11 predictors, on average, compared with 7 for GEL.

The superior performance of GEL is seen most dramatically with respect to group selection, where it outperformed the other four methods by a considerable margin. Notably, in the 20 predictors per group setting, GEL achieved near-perfect group selection accuracy, committing zero misclassifications in 92% of the simulations. In contrast, no other method was able to achieve perfect classification at the group level in more than 20% of the simulations.

In summary, for all three measures (estimation accuracy, group selection, and individual variable selection), GEL clearly outperforms the four other methods, although its performance was similar to composite MCP for small groups.

3.2. *Varying the Degree of Grouping*

The generating model for each simulation in this section contains eight nonzero coefficients, distributed into a varying number of groups, $\{8, 4, 2, 1\}$. At the one extreme, all eight coefficients are in different groups and the knowledge of which group the predictors belong to is not useful. At the other extreme, all eight coefficients belong to the same group, and the grouping information is highly useful. Here, we fix J, the number of groups, at 100, each containing 10 elements (i.e., the total number of predictors is 1000). Different versions of the GEL, with τ varying among $\{0.1, 0.2, 0.3, 0.5\}$, were fit to each simulated data set in order to examine the empirical role of the τ parameter.

The results of the simulation are presented in Figure 4. Along the horizontal axis, we plot the number of non-null groups, although the axis may be informally thought of as the "degree of grouping," with the far left being "no grouping" and the far right denoting a rather heavy degree of grouping. Varying this axis has no impact whatsoever on the MCP estimates, which do not utilize grouping information in any way. The degree of grouping, however, has a noticeable impact on the GEL estimates.

It is worth noting that even in the case where the solutions are not grouped in any way, the estimation accuracy of GEL is fairly similar to that of MCP. In particular, for $\tau = 0.3$, the GEL is only 5% less efficient than MCP. Thus, GEL is robust to the assumption of grouping—the price one pays for an incorrect assumption of grouping, or for classification error in grouping, is rather small (although it can be larger when the signal-to-noise ratio is high, see supplementary materials).

When grouping information is valuable, however, the GEL leverages this information to considerable effect. Indeed, at the highest level of grouping considered here, the estimation accuracy of GEL is five times better than that of MCP. This finding is fairly similar across the values of τ considered here, although of course as $\tau \to 0$, the GEL approaches the lasso (Proposition 2), and the advantage is lost.

A similar result holds, at least for $\tau = 0.3$ and $\tau = 0.5$, with respect to variable selection. These methods are comparable to MCP at low levels of grouping, but clearly outperform MCP at the highest degree of grouping: GEL ($\tau = 0.3$) committed only 4 misclassification errors on average, compared with 12 for MCP. GEL with $\tau = 0.2$ and $\tau = 0.1$, on the other hand, are considerably worse at the selection of individual variables than MCP is. The reason for this is that these estimators are becoming lasso-like at small values of τ, and the lasso is considerably worse at variable selection than is MCP. The lasso averaged over 30 misclassifications per simulation; it is not shown on the graph for the sake of clarity, but if it were, would be represented in the middle panel by a horizontal line near 30. This is also why increasing τ decreases misclassifications even in the absence of grouping.

Similar trends hold for group selection, although GEL even more dramatically outperforms MCP in this setting. At the highest degree of grouping, GEL ($\tau = 0.3$) commits only two misclassifications at the group level (out of 100 groups), while MCP averaged 8 misclassifications. Again, for lower values of τ, GEL begins to exhibit similar selection properties to that of the lasso, which here commits an average of 24 misclassi-

fications (once again, not shown on the graph for the sake of clarity).

These simulation results demonstrate a robustness to the selection of τ not apparent from the derivation of the GEL in Section 2.2. On the one hand, this may be thought of as a drawback, in that the method does not allow the analyst to modify the strength of coupling but leave convexity unaltered. In practice, however, this makes the method much simpler to apply, in that the analyst does not need to worry a great deal about extra tuning parameters. Indeed, the GEL with $\tau = 0.3$ achieves a nice balance across various degrees of grouping, exhibiting similar performance to that of MCP (arguably the state of the art in terms of non-grouped variable selection methods) in the ungrouped case, while making effective use of grouping information when it is useful.

We find $\tau \approx 1/3$ to be broadly successful as a default value, with relatively diminishing returns at larger τ values. Certainly, Figure 4 shows that $\tau = 0.5$ is capable of offering advantages over $\tau = 0.3$, particularly in terms of group selection accuracy. However, as shown in Proposition 1, increasing τ decreases the convexity of the objective function and risks the possibility of the solution to (2) exhibiting multiple local minima. In practice, we found $\tau = 1/2$ to occasionally run into difficulty with convergence, while these problems were rare for the GEL with $\tau = 1/3$.

4. Application to Association Studies Involving Rare Variants

An important potential application of bi-level selection in general, and the GEL in particular, is the problem of identifying rare variants associated with disease in genetic association studies. Briefly, the idea is that many genetic variants of a given gene may exist. Each individual variant may be rare, but collectively, the group of variants could play a significant role in phenotypic variation in the population. A typical analysis that tests each variant separately, however, would have low power to detect associations involving any of these variants (see Li and Leal, 2008; Bansal et al., 2010, for more extensive discussion of these issues).

Currently, most methods for addressing this problem involve one of the following two approaches: "collapsing" the rare variants into a single measurement, or carrying out an omnibus multivariable/multivariate test of the joint null hypothesis that none of the variants in a gene have an effect. Each approach has its limitations, however, and neither approach provides any insight into the selection of individual variants. In particular, neither approach can address the question of choosing between individual variants in different genes.

An alternative approach is to use group penalization. Here, each variant is a predictor and the variants are grouped according to the gene they belong to. Such a model naturally pools information across variants in a group, while avoiding the simplistic assumption (made by the "collapsing" approach) that each variant has exactly the same effect on the phenotypic outcome. Furthermore, the GEL performs bi-level selection, providing insight into which individual variants—as well as which genes—are responsible for phenotypic variation.

To test this approach on real(istic) data, we analyzed the data set from the 2010 Genetic Analysis Workshop (GAW). The data set contains real exon sequencing data from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010) on 697 unrelated individuals and 24,487 genetic variants, grouped into 3205 genes (in the notation of this paper, $n = 697$, $p = 24,487$, and $J = 3,205$). Two hundred independent sets of quantitative phenotypes were simulated by the organizers of the workshop according to a plausible genetic model of variant-disease association; the data and simulation design are described in greater detail in Almasy et al. (2011).

We compared the GEL with the following methods, some of which perform only variant-level selection and others performing gene-level selection: the lasso, which performs variant-level selection; the "Univariate" approach, which performs variant-level testing based on basic univariate tests of correlation; the group lasso, which performs gene-level selection; the "Collapse" approach, which performs gene-level testing based on adding together the variants in that gene; and the "Multivariate" approach, which performs gene-level testing based on an omnibus F-test in the linear regression model with all variants for that gene included as predictors. The composite MCP and SGL, like the GEL, perform bi-level selection and are included in both gene-level and variant-level comparisons.

To make the comparison equivalent between variable-selection approaches and hypothesis-testing approaches, each method was allowed to select the same number of variants/genes. Since there were 39 causal variants in the generating model, the variant-level approaches were allowed to select 39 variants (the first 39 coefficients to enter the model for the variable-selection approaches, the most significant 39 variants for the hypothesis testing approach). Likewise, as there were 9 causal genes, each gene-level approach was allowed to select 9 genes. The bi-level selection approaches were included in both comparisons, although the final models appearing in each comparison are slightly different depending on whether the gene-level stopping rule or the variant-level stopping rule was used. Methods were compared according to the accuracy of these selections. In practice, of course one would not know the true number of causal variants/genes. However, it is often the case that researchers only have the resources to follow-up on a limited number of findings. For this reason, it is not uncommon in practice to select the top, say, 10 or 25 features for further follow-up. In summary, although the exact number (9, 39) of selections is artificial, the design of the study is a fair comparison of realistic criteria. Information on the performance of cross-validation for penalized regression methods is available in the supplementary materials.

The results of the analyses, averaged over the 200 simulation replications, are in presented for the variant-selection methods and gene-selection methods in Tables 1 and 2, respectively. As shown in Table 1, the univariate and lasso methods do not incorporate grouping information into the selection process; consequently, the 39 variants they select are scattered across 30+ genes. Clearly, since the causal variants are concentrated in 9 genes, most of these selections are necessarily false positives. Indeed, only ≈ 4 of the 39 variants selected by these methods was in fact causal.

The composite MCP, although it in principle uses group information, fails to do so effectively in this problem, performing rather similarly to the univariate and lasso methods. As shown in Figure 2, the degree of coupling in the com-

Table 1
GAW results for methods that carry out variant-level selection; each method selected its top 39 variants

	Number of genes selected	Causal variants selected
Univariate	30.1	3.9
Lasso	35.5	4.3
MCP	36.7	3.3
SGL	23.6	5.1
Composite	36.0	3.9
GEL	6.3	11.3

posite MCP diminishes with group size. For these data, the median group size among the 9 causative genes was 10, with two genes (FLT1 and HIF3A) containing over 20 variants. Given these group sizes, it is perhaps unsurprising that the composite MCP performed similarly to lasso.

Unlike the other methods, the SGL and the GEL successfully incorporate grouping information to increase the accuracy of variable selection. However, GEL was far more effective than SGL at utilizing group information. Its 39 selections are concentrated into just ≈ 6 groups, roughly matching the actual degree of grouping in the generating model. By leveraging this group information, GEL is able to achieve much greater accuracy with its selection of variants, correctly identifying on average 2–3 times the number of causal variants identified by the other methods.

Table 2 presents the accuracy of the gene-level selection methods. The most striking result is the extremely poor performance of the group lasso; in 178 out of the 200 simulations, it failed to select even a single causative gene. The reason for this would seem to be the fact that if the group lasso selects a gene, each of its variants enters the model. Even in the causative genes, however, most variants are not causally related to the outcome; for example, only 3 out of the 21 HIF3A variants have a nonzero coefficient in the generating model. The group lasso displays a strong bias here toward selecting single-variant groups, reflected in the fact that even though it selected 9 genes and all variants in those genes, this amounted to only 9.4 variants. The SGL performs somewhat better than the group lasso, but still much worse than gene-level testing.

As in Table 1 GEL emerges as the best of the methods being compared. It selects, on average, the largest number of causal genes, although the difference is not as striking as in the variant-level comparison, where GEL outperformed hypothesis testing by 286%: here, GEL outperforms the "collapse" testing by 23% and "multivariate" testing by 12%. An added benefit of GEL is that, unlike the two hypothesis test approaches, GEL is able to narrow down the list of important variants to 45 (quite close to the true number of 39), while the multivariate and collapse approaches produce 99 and 146, respectively. Furthermore, unlike a hypothesis testing approach, GEL produces estimates of effect sizes and a predictive model.

Table 2
GAW results for methods that carry out gene-level selection; each method selected its top 9 genes

	Number of variants selected	Causal genes selected
Collapse	146.5	1.3
Multivariate	98.8	1.4
Group lasso	9.4	0.1
SGL	14.9	0.4
Composite	10.9	1.5
GEL	45.4	1.6

5. Discussion

In this paper, we have introduced the notion of group exponential penalties as a method for leveraging grouping information among features in order to achieve bi-level variable selection—the selection of important groups as well as the important individual predictors in those groups. This idea is both theoretically attractive and practically useful. In particular, we demonstrate a clear potential for group exponential penalties as a method for the analysis of rare variants in genetic association studies. The approach is computationally efficient, scales up to high dimensions, and publicly available via the `grpreg` package. Although we have concentrated on linear regression in this paper, our implementation of GEL for logistic regression is also implemented in `grpreg`, and the idea extends readily to other regression models such as Poisson and proportional hazards models.

A potential further extension of this idea is the "group exponential MCP" estimator alluded to at the end of Section 2.2, in which the inner L_1 norm is replaced by the minimax concave penalty. The L_1 norm introduces a bias toward zero that may diminish estimation accuracy. This bias is alleviated by the outer exponential penalty, but not completely eliminated. A group exponential MCP estimator may be able to improve upon this result.

At the same time, however, the GEL offers a rather elegant simplicity in that there seems to be little need to tune τ, and thereby concern oneself only with λ during analysis; this is a considerable asset in practice. Indeed, the GEL with $\tau = 1/3$ exhibits remarkably robust improvements in performance over existing methods in terms of both estimation and selection accuracy across a range of group sizes and degrees of grouping.

6. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2 and 3 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The author would like to thank the Associate Editor and referee for helpful comments and suggestions that improved the quality of this manuscript.

References

Almasy, L., Dyer, T., Peralta, J., Kent, J., Charlesworth, J., Curran, J., and Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings* **5**, S2.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11**, 773–785.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface* **2**, 369–380.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* **5**, 232–253.

Donoho, D. and Johnstone, J. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

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

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**, 481–499.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339–355.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *American Journal of Human Genetics* **83**, 311–321.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.

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

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics* **2**, 224–244.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38**, 894–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of The American Statistical Association* **101**, 1418–1429.

Received March 2014. *Revised November* 2014.
Accepted February 2015.