Exploiting Covariate Similarity in Sparse Regression via the Pairwise Elastic Net

Alexander Lorbert*, **David Eis***, **Victoria Kostina***, **David M. Blei**†, **Peter J. Ramadge***
*Dept. of Electrical Engineering, †Dept. of Computer Science
Princeton University, Princeton, NJ 08544

Abstract

A new approach to regression regularization called the Pairwise Elastic Net is proposed. Like the Elastic Net, it simultaneously performs automatic variable selection and continuous shrinkage. In addition, the Pairwise Elastic Net encourages the grouping of strongly correlated predictors based on a pairwise similarity measure. We give examples of how the approach can be used to achieve the objectives of Ridge regression, the Lasso, the Elastic Net, and Group Lasso. Finally, we present a coordinate descent algorithm to solve the Pairwise Elastic Net.

1 INTRODUCTION

In a standard linear regression problem we are given n measurements of a p-dimensional input vector along with the corresponding responses and we wish to estimate the weights of a linear model that optimize both accuracy and parsimony. Accuracy is typically measured by least-squared error. Parsimony may be measured by the number of non-zero weights required by the model, although for computational reasons this is typically relaxed to a convex penalty (ℓ_1).

A significant issue in estimating the weights arises when the regressors, or groups of regressors, are highly correlated or clustered. The Lasso (Tibshirani, 1996) (ℓ_1-regularization) will generally select a single representative from each cluster and ignore other cluster members. This leads to parsimonious (sparse) solutions, but the model misses the important cluster structure in the data. Indeed, there is no qualitative reason to choose one feature over another among a cluster of highly correlated regressors. On the other hand, Ridge regression (Hoerl & Kennard, 1970) (ℓ_2-regularization) will generally "group" all clustered features, i.e., each regressor is assigned a weight similar to others in its cluster. This is a safe approach when all features are generally deemed relevant, but parsimony is not achieved, and once again, important structure in the data has not been identified in the model.

A method that fuses both the Lasso and Ridge is the Elastic Net (Zou & Hastie, 2005). The Elastic Net encourages both sparsity and grouping by forming a convex combination of the Lasso and Ridge regularization governed by a selectable parameter. Furthermore, unlike the Lasso, the Elastic Net can yield a sparse estimate with more than n non-zero weights (Efron et al., 2004). One can view the Elastic Net as placing a *global* tradeoff between sparsity (ℓ_1) and grouping (ℓ_2).

The sparsity/grouping tradeoff was also addressed with the Group Lasso (Yuan & Lin, 2006). Given prior knowledge of how to partition the features into clusters or groups, the Group Lasso produces a sparse cluster solution. Precise prior knowledge of groups is a limiting requirement. Additionally, the Group Lasso assumes each feature is a member of one group only. In practice, one may desire a feature to belong to multiple groups.

We introduce an approach for establishing local, or pairwise, tradeoffs using a user definable measure of similarity between regressors. This allows for a more adaptive grouping than that permitted by the Elastic Net. We call our method the *Pairwise Elastic Net* (PEN). To motivate the idea of leveraging local sparsity/grouping tradeoffs, consider the following two feature correlation matrices:

$$R_1 = \begin{pmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{pmatrix} \qquad R_2 = \begin{pmatrix} 1.0 & 0.9 & 0.0 \\ 0.9 & 1.0 & 0.4 \\ 0.0 & 0.4 & 1.0 \end{pmatrix} \tag{1}$$

Matrix R_1 depicts features that have the same pairwise correlation. Such a global relationship would mo-

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

tivate use of the Elastic Net. In contrast, in R_2, features 1 and 2 are highly correlated - an argument for Ridge; features 1 and 3 are orthogonal - an argument for Lasso; features 2 and 3 are slightly correlated, suggesting the Elastic Net. However, assigning a single global sparsity/grouping tradeoff, as required in the Elastic Net, ignores the local information available in the data. The Pairwise Elastic Net leverages local sparsity/grouping tradeoffs, thereby allowing more flexibility than the Elastic Net. This can match up regularization to evident structure in the data.

In summary, our main contribution is to put forth the proposal of the Pairwise Elastic Net (PEN), an approach for establishing local, or pairwise, tradeoffs in regression regularization using a user-definable measure of regressor similarity. We give some examples how this framework encompasses many related ideas for regression regularization and derive a result on its ability to "group" the estimated weights of similar regressors. We then provide a coordinate descent algorithm to efficiently solve the PEN regression problem. Finally, we test the PEN on real-world and simulated datasets.

The remainder of the paper is organized as follows. In §2 we introduce the PEN, and give some insights on its attributes and flexibility. In §3, we prove that the Pairwise Elastic Net assigns similar regression coefficients to similar features, i.e., exhibits the grouping effect. The coordinate descent algorithm is described in §4, and §5 discusses the rescaling of the Pairwise Elastic Net solution similar to that present in the Elastic Net. §6 presents examples from simulated and real-world data.

2 PAIRWISE ELASTIC NET

Consider a linear model:

$$y = X\beta_* + \varepsilon \tag{2}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix, $y \in \mathbb{R}^n$ is the response vector, $\beta_* \in \mathbb{R}^p$ are the unknown weights, and $\varepsilon \in \mathbb{R}^n$ is a zero-mean i.i.d. Gaussian noise vector. Let X_i denote the i-th column of X. We call X_i the i-th *feature* or *regressor* vector. Without loss of generality, we assume that our model in (2) is standardized:

$$\mathbf{1}^T y = 0 \quad \mathbf{1}^T X_i = 0 \quad X_i^T X_i = 1 \tag{3}$$

where $i = 1, \ldots, p$ and $\mathbf{1}$ denotes the vector of all 1's.

The generic penalization method finds an estimate $\hat{\beta}$ of β_* by solving the following minimization problem:

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2 + \eta J(\beta), \tag{4}$$

where $J(\beta)$ is a nonnegative valued penalty function, and η is a nonnegative complexity parameter. For ordinary least squares $J(\beta) = 0$, for Ridge regression $J(\beta) = \|\beta\|_2^2$, and for the Lasso $J(\beta) = \|\beta\|_1$. The Elastic Net uses an additional parameter $\alpha \in [0,1]$ and has $J(\beta) = \alpha\|\beta\|_2^2 + (1-\alpha)\|\beta\|_1$.

We would like to find a convex function $J(\beta)$ that simultaneously encourages sparsity and grouping based on *pairwise* feature similarity, i.e., if features X_i and X_j are similar, then $\hat{\beta}_i$ and $\hat{\beta}_j$ should have similar values. To this end, we introduce a *feature similarity matrix* $R \in \mathbb{R}^{p\times p}$ where R_{ij} is a measure of the "similarity" of X_i and X_j. Examples of similarity measures are the absolute sample correlation, $R_{ij} = |X_i^T X_j|$ and the midpoint metric $R_{ij} = \|X_i + X_j\|/2 = (1/\sqrt{2})\sqrt{1 + X_i^T X_j}$. More interesting examples include the positive semidefinite Gaussian kernel $R_{ij} = \exp(-\|X_i - X_j\|^2/\sigma^2)$. The selection of R can also be used to incorporate prior knowledge. For example: (a) perfectly similar features $R = \mathbf{1}\mathbf{1}^T$; (b) perfectly dissimilar features $R = I$; and (c) known group structure: $R_{ij} = 1$ if features i and j belong to the same group and is 0 otherwise (see Fig. 1(c)). All of the above examples fall into the following general class. Select a $\rho\colon \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ as a measure of similarity. It is natural to ask that for $z_1, z_2 \in \mathbb{R}^n$:

$$0 \le \rho(z_1, z_2) = \rho(z_2, z_1) \le 1 \quad \rho(z_1, z_1) = 1 \tag{5}$$

then set $R = [\rho(X_i, X_j)]$. In this case, R is symmetric with nonnegative entries, ones down the diagonal, and off-diagonal entries bounded above by 1. It is also possible to select R so that, in addition to the above, it is positive semidefinite. Moreover, we might weight, threshold, or smooth the similarities to incorporate additional attributes.

Before proposing the Pairwise Elastic Net penalty, we introduce some notation and basic results. Let $|\beta|$ denote the vector with $|\beta|_i = |\beta_i|$, $i = 1, \ldots, p$. Observe that:

$$\begin{aligned} \|\beta\|_2^2 &= |\beta|^T|\beta| \\ \|\beta\|_1 &= |\beta|^T\mathbf{1} = \mathbf{1}^T|\beta| \\ \|\beta\|_1^2 &= |\beta|^T\mathbf{1}\mathbf{1}^T|\beta| \end{aligned} \tag{6}$$

with $\mathbf{1}\mathbf{1}^T$ being the matrix of all ones. We also note the following important result, which will be used later.

Lemma 2.1. *Let $P \in \mathbb{R}^{p\times p}$ be symmetric. Then*

$$J(\beta) = |\beta|^T P\, |\beta| \tag{7}$$

is convex if and only if P has nonnegative entries and is a positive semidefinite matrix.

Proof. First assume that P is nonnegative and positive semidefinite. We show that $J(\delta\alpha+(1-\delta)\beta) \leq \delta J(\alpha)+(1-\delta)J(\beta)$ for $\delta \in [0,1]$ and $\alpha, \beta \in \mathbb{R}^p$. We have:

$$\begin{aligned}
&J(\delta\alpha + (1-\delta)\beta)\\
&= \sum_{i,j} P_{ij}|\delta\alpha_i + (1-\delta)\beta_i| \cdot |\delta\alpha_j + (1-\delta)\beta_j|\\
&\leq \sum_{i,j} P_{ij}(\delta^2|\alpha_i| \cdot |\alpha_j| + (1-\delta)^2|\beta_i| \cdot |\beta_j|\\
&\quad + 2\delta(1-\delta)|\alpha_i| \cdot |\beta_j|) \quad \text{since } P_{ij} \geq 0\\
&= \delta^2|\alpha|^T P|\alpha| + (1-\delta)^2|\beta|^T P|\beta| + 2\delta(1-\delta)|\alpha|^T P|\beta|\\
&\leq \delta|\alpha|^T P|\alpha| + (1-\delta)|\beta|^T P|\beta|
\end{aligned}$$

In the last line, by the PSD of P: $(x-y)^T P(x-y) \geq 0$, i.e., $2x^T Py \leq x^T Px + y^T Py$. Hence $2\delta(1-\delta)|\alpha|^T P|\beta| \leq \delta(1-\delta)(|\alpha|^T P|\alpha| + |\beta|^T P|\beta|)$.

Now, assume that J is convex. First we show that the diagonal elements of P must be nonnegative. Let $\alpha = e_k$ and $\beta = -e_k$, where e_k is the kth standard basis vector of $\mathbb{R}^p$. By the convexity of J, we have

$$\begin{aligned}
\left|\tfrac{\alpha+\beta}{2}\right|^T P \left|\tfrac{\alpha+\beta}{2}\right| &\leq \frac{1}{2}|\alpha|^T P|\alpha| + \frac{1}{2}|\beta|^T P|\beta|\\
0 &\leq P_{kk}
\end{aligned}$$

for $k = 1, 2, \ldots, p$. Thus, all diagonal elements must be nonnegative. Starting with nonnegative diagonal elements, we prove that all off-diagonal elements must be nonnegative. Let $\alpha = e_i + \tau e_j$ and $\beta = e_i - \tau e_j$ for any (i,j) with $i \neq j$, where we require $\tau > 0$. This yields $|\alpha| = |\beta| = \alpha$ and

$$\begin{aligned}
J(\alpha) = J(\beta) &= (e_i + \tau e_j)^T P(e_i + \tau e_j)\\
&= e_i^T Pe_i + \tau e_i^T Pe_j + \tau e_j^T Pe_i + \tau^2 e_j^T Pe_j\\
&= P_{ii} + 2\tau P_{ij} + \tau^2 P_{jj}
\end{aligned}$$

We also have $J((\alpha+\beta)/2) = J(e_i) = P_{ii}$. By convexity, for all positive τ we have

$$\begin{aligned}
\left|\tfrac{\alpha+\beta}{2}\right|^T P \left|\tfrac{\alpha+\beta}{2}\right| &\leq \frac{1}{2}|\alpha|^T P|\alpha| + \frac{1}{2}|\beta|^T P|\beta|\\
P_{ii} &\leq P_{ii} + 2\tau P_{ij} + \tau^2 P_{jj}\\
0 &\leq 2P_{ij} + \tau P_{jj}
\end{aligned}$$

First, if $P_{jj} = 0$ then $P_{ij} \geq 0$. Now suppose $P_{jj} > 0$ and assume that $P_{ij} < 0$. If we let $\tau = -P_{ij}/P_{jj}$, which is strictly greater than zero, then $0 \leq 2P_{ij} - P_{ij} = P_{ij}$, resulting in a contradiction. Hence, all elements of P are nonnegative.

Finally, we show that P is positive semidefinite. Let $\mu \in \mathbb{R}$ denote the minimum eigenvalue of P with corresponding unit-norm eigenvector $u \in \mathbb{R}^p$. We set $\alpha = \mathbf{1}+\tau u$ and $\beta = \mathbf{1}-\tau u$ where $0 < \tau \leq 1/\max_i |u_i|$. By construction, α and β are nonnegative-valued vectors so $|\alpha| = \alpha$ and $|\beta| = \beta$. Since J is convex,

$$\begin{aligned}
\left|\tfrac{\alpha+\beta}{2}\right|^T P \left|\tfrac{\alpha+\beta}{2}\right| &\leq \frac{1}{2}|\alpha|^T P|\alpha| + \frac{1}{2}|\beta|^T P|\beta|\\
\mathbf{1}^T P\mathbf{1} &\leq \mathbf{1}^T P\mathbf{1} + \tau^2 u^T Pu\\
0 &\leq \tau^2\mu
\end{aligned}$$

implying that the minimum eigenvalue of P must be nonnegative. Thus, P must be PSD. □

We now introduce the pairwise Elastic Net:

$$\hat{\beta} = \arg\min_\beta \|y - X\beta\|_2^2 + \eta\, |\beta|^T P\, |\beta|, \qquad (8)$$

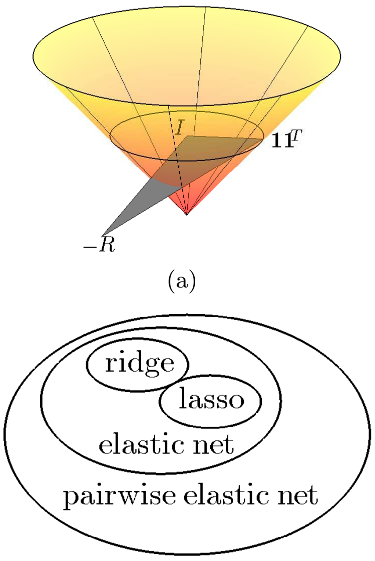
where P is a symmetric, PSD matrix with nonnegative entries. We presently discuss some possible ways of choosing P. The standard penalty matrices I and $\mathbf{1}\mathbf{1}^T$ lie in the cone of PSD matrices with nonnegative entries in $\mathbb{R}^p$. If we apply a global tradeoff between an ℓ_1-squared and ℓ_2 penalties, à la the Elastic Net, then we form a convex combination $P_\alpha = (1-\alpha)I + \alpha\mathbf{1}\mathbf{1}^T$. P_α lies on the line between I and $\mathbf{1}\mathbf{1}^T$ within the cone. We now bring in the similarity matrix R and consider points in the cone of PSD matrices formed from I, $\mathbf{1}\mathbf{1}^T$ and R. For example, starting from any point on the above line segment, heading in the direction of $-R$ allows us to incorporate pairwise tradeoffs, e.g., $P_{\alpha,\mu} = (1-\alpha)I + \alpha\mathbf{1}\mathbf{1}^T - \mu R$ (see Fig. 1(a)). To use Lemma 2.1, the resulting matrix $P_{\alpha,\mu}$ must stay in the cone of PSD matrices and have nonnegative entries. As a second example, select a similarity matrix R that is PSD and nonnegative; then form $P = \alpha_1 I + \alpha_2\mathbf{1}\mathbf{1}^T + \alpha_3 R$, where $\alpha_i \geq 0$, $i = 1, 2, 3$, and $\sum_{i=1}^3 \alpha_i = 1$. In this case, P has nonnegative entries and is PSD. As a third example, consider

$$P = I + \mathbf{1}\mathbf{1}^T - R, \qquad (9)$$

with $R_{ij} = \rho(X_i, X_j)$ as outlined above. This yields

$$\begin{aligned}
J(\beta) &= |\beta|^T(I + \mathbf{1}\mathbf{1}^T - R)|\beta| \qquad (10)\\
&= \|\beta\|_2^2 + \|\beta\|_1^2 - |\beta|^T R|\beta|.
\end{aligned}$$

The third term in (10) represents a tradeoff between ℓ_1-squared and ℓ_2 regularization in which we cut back the ℓ_1-squared penalty when the similarity between the corresponding features is high. For perfectly similar features, $R = \mathbf{1}\mathbf{1}^T$, (10) reduces to the Ridge penalty $J(\beta) = \|\beta\|_2^2$ and for perfectly dissimilar features, $R = I$, it reduces to the ℓ_1-squared penalty $J(\beta) = \|\beta\|_1^2$, which is equivalent to the Lasso (Theorem 2.3). To illustrate the "pairwise" property, suppose that $\beta_k = 0$ for $k \neq i, j$. Then the regularization function simplifies to $J(\beta) = R_{ij}(\beta_i^2 + \beta_j^2) + (1-R_{ij})(|\beta_i| + |\beta_j|)^2$ which is a convex combination of ℓ_1-squared and ℓ_2 penalties. Thus, greater similarity leads to less ℓ_1-squared



(a)

(b)

$$R = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

(c)

Figure 1: (a) Geometry of ℓ_1-squared/ℓ_2 Tradeoff. (b) Venn Diagram of β-estimators. (c) A Similarity Matrix Exhibiting Group Structure.

penalization, and vice versa. This intuitive result can be restated as follows: when two features are similar, we take a Ridge approach; when two features are dissimilar, we take a Lasso approach.

The matrix (9) is not always PSD, which is required to apply Lemma 2.1. But in this case a shrinkage parameter can be used to ensure that P is PSD. There are several ways this can be done. Here we consider

$$P_\theta^S = \theta I + (1-\theta)P = I + (1-\theta)\mathbf{11}^T - (1-\theta)R \quad (11)$$

In this case we have the following lemma.

Lemma 2.2. *Let* $P \in \mathbb{R}^{n\times n}$ *be symmetric,* $\tau = -\min\{0, \lambda_{min}(P)\}$*, and*

$$P_\theta^S = \theta I + (1-\theta)P, \quad (12)$$

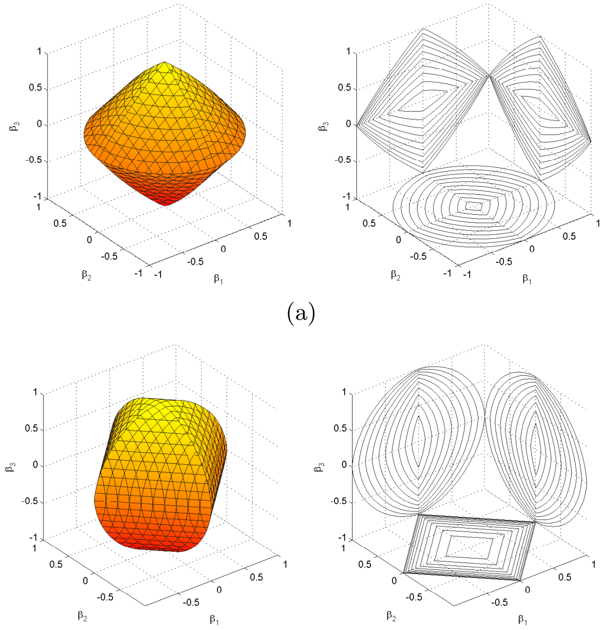
Then for $\frac{\tau}{1+\tau} \leq \theta \leq 1$*,* P_θ^S *is* PSD.

Proof. $\lambda_{min}(\theta I + (1-\theta)P) = \theta + (1-\theta)\lambda_{min}(P) \geq \theta - (1-\theta)\tau \geq \frac{\tau}{1+\tau} - \frac{1}{1+\tau}\tau = 0.$ □

Finding the minimum eigenvalue of a symmetric matrix is a convex minimization problem (Boyd & Vandenberghe, 2004), so, if required, the shrinkage proposed above is feasible. Note that if P has nonnegative entries then so does P_θ^S. Hence for P given in (9), P_θ^S has nonnegative entries and Lemma 2.1 guarantees that (4) with $J(\beta) = |\beta|^T P_\theta^S |\beta|$ is a convex optimization problem. We emphasize, however, that there are ways to select the similarity matrix R so that convex combinations of I, $\mathbf{11}^T$ and R are nonnegative and PSD.

2.1 SELECTING R

To illustrate some of the PEN's basic attributes and to demonstrate its versatility, we consider below several instances of how one might exploit covariate structure. In doing so we relate the PEN to both the Elastic Net and Group Lasso.



(a)

(b)

Figure 2: Penalty Surfaces and Their Contours.

First, to illustrate the PEN in a visualizable context, consider three features with their pairwise similarity measure. Two features are highly similar: $R \in \mathbb{R}^3$ with $R_{12} = 0.9$, $R_{13} = 0$, and $R_{23} = 0.1$. One would like a penalty surface in which, pairwise, features 1 and 2 have a Ridge-like penalty, while the other feature pairs have a Lasso-like penalty. P defined in (9) is not PSD. The PEN surface using P_θ^S with $\theta = 0.23$ is plotted in Fig. 2(a) with the projected contours shown on each coordinate pair plane. Note the Ridge-like contours on the 1-2 plane and the Lasso-like contours on the 1-3 and 2-3 planes. As a second illustration, suppose two dissimilar features are highly similar with the third: $R \in \mathbb{R}^3$ with $R_{12} = 0.0, R_{13} = 0.9, R_{23} = 0.9$. In this case, one would like a penalty surface in which, pairwise, features 1 and 2 have a Lasso-like penalty, while the other feature pairs have a Ridge-like penalty. The PEN surface using P is plotted in Fig. 2(b) with the projected contours shown on each coordinate pair plane. Note the Lasso-like contours on the 1-2 plane and the Ridge-like contours on the 1-3 and 2-3 planes.

To make a connection with the Elastic Net, suppose the features possess a global pairwise similarity: $R_{ii} = 1$ and $R_{ij} = \sigma$ $(i \neq j)$ with $0 \leq \sigma \leq 1$. Then for $\mu = 1$, $P_\sigma = I + \mathbf{1}\mathbf{1}^T - R = \sigma I + (1-\sigma)\mathbf{1}\mathbf{1}^T$ with eigenvalues $\sigma + (1-\sigma)p$ (with multiplicity 1) and σ (with multiplicity $p-1$). Thus P_σ is symmetric, positive definite, has positive entries and the PEN penalty is:

$$\begin{aligned} J_\sigma(\beta) = |\beta|^T P_\sigma |\beta| &= \sigma \sum_{i=1}^{p} \beta_i^2 + (1-\sigma)\sum_{i,j}^{p} |\beta_i \beta_j| \\ &= \sigma \|\beta\|_2^2 + (1-\sigma)\|\beta\|_1^2 , \end{aligned}$$

Now, we define two problems,

$$\begin{aligned} &P1 : \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2 \\ &P2 : \arg\min_{\beta} \|y - X\beta\|_2^2 + \eta J_\sigma(\beta), \end{aligned}$$

where $P1$ is Elastic Net. We assume that λ_1 and λ_2 are not both equal to zero. Then Theorem 2.3 shows that $P1$ and $P2$ are equivalent.

Theorem 2.3. *Fix $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$. If $\widetilde{\beta} \neq \mathbf{0}$ is a solution of $P1$, then it solves $P2$ with $\eta = (2\lambda_2\|\widetilde{\beta}\|_1 + \lambda_1)/(2\|\widetilde{\beta}\|_1)$ and $\sigma = 1 - \lambda_1/(2\lambda_2\|\widetilde{\beta}\|_1 + \lambda_1)$. Conversely, if $\widetilde{\beta}$ is a solution of $P2$ then it solves $P1$ with $\lambda_1 = 2\eta(1-\sigma)\|\widetilde{\beta}\|_1$ and $\lambda_2 = \eta\sigma$.*

Proof. The KKT equations for $P1$ and $P2$, respectively, are

$$\begin{aligned} -2X_j^T y + 2X_j^T X\beta_j + \lambda_1 \mathrm{sgn}(\beta_j) + 2\lambda_2\beta_j = 0 \\ -2X_j^T y + 2X_j^T X\beta_j + 2\eta(1-\sigma)\mathrm{sgn}(\beta_j)\|\beta\|_1 \\ + 2\eta\sigma\beta_j = 0 \end{aligned}$$

for $j = 1, \ldots, p$. For $\lambda_1 = 2\eta(1-\sigma)\|\widetilde{\beta}\|_1$ and $\lambda_2 = \eta\sigma$ the above two series of equations are equivalent when $\beta = \widetilde{\beta}$. Thus, if $\widetilde{\beta}$ is a solution to one, it is a solution to the other. Solving for η and σ in terms of λ_1 and λ_2 yields the relationships in the theorem. □

We note that P_σ can correspond to the similarity matrix $R = (1-\sigma)I + \sigma\mathbf{1}\mathbf{1}^T$. Furthermore, σ defined in terms of λ_1 and λ_2 must lie in the interval $[0, 1]$. When $\sigma = 1$, we have Ridge regression, and $\sigma = 0$ gives an ℓ_1-squared regularizer, which is equivalent to the Lasso.

Finally, we draw a connection with Group Lasso. Suppose the features form m known disjoint groups. In this case, define group indicator vectors $g_k \in \mathbb{R}^p$, $k = 1 \ldots m$, with $g_k(i) = 1$ if and only if feature i belongs to group k. Let $p_k = \mathbf{1}^T g_k$ denote the size of group k and set $G = [g_1, \ldots, g_m]$. Note that since the groups are exclusive and every feature must be in one group, $G\mathbf{1} = \mathbf{1}$. For the similarity matrix we use $R = G\Sigma^{-1}G^T$ where $\Sigma = \mathrm{diag}(p_1, \ldots, p_m)$. Without the scaling by Σ^{-1} this would yield a matrix like that shown in Fig. 1(c). The scaling divides the entries of R for each group by the group size. Now combine R with I and $\mathbf{1}\mathbf{1}^T$ as follows to form the penalty matrix P:

$$P = (1-\alpha)I + \alpha\mathbf{1}\mathbf{1}^T - (1-\alpha)R \qquad (13)$$

Let $x \in \mathbb{R}^p$ and $w = G^T x \in \mathbb{R}^m$. Then, by straightforward algebraic expansion, we find

$$\begin{aligned} x^T P x &= x^T\left((1-\alpha)(I - G\Sigma^{-1}G^T) + \alpha G\mathbf{1}\mathbf{1}^T G^T\right)x \\ &= (1-\alpha)\sum_{k=1}^{m}\sum_{g_k(i)=1}(x_i - \bar{x}_k)^2 + \alpha\Big(\sum_k w_k\Big)^2 \end{aligned}$$

where $\bar{x}_k = 1/p_k \sum_{g_k(i)=1} x_i$ is the mean of x over group k. Hence P is positive definite for all $\alpha \in [0, 1]$. For positive-valued P we require $\alpha > 1/(1 + \min_k p_k)$, which is derived from the structure of R. Now setting $x = |\beta|$ yields

$$|\beta|^T P|\beta| = (1-\alpha)\sum_{k=1}^{m}\sum_{g_k(i)=1}(|\beta_i| - \overline{|\beta|}_k)^2 + \alpha\|w\|_1^2$$

where $\overline{|\beta|}_k = 1/p_k \sum_{g_k(i)=1}|\beta_i|$ is the mean absolute weight assigned over group k. The first term measures the variation of coefficients' absolute value about the mean in each group. This penalty encourages uniformity of absolute weights within groups. The second term is an ℓ_1^2 penalty on the distribution of weights across groups. This encourages sparsity of group selection and α controls the tradeoff between these objectives. Thus we have the PEN performing a regularization akin to the Group Lasso.

It is also possible to select R to encode a "soft grouping" in which each regressor is assigned a probability mass function over m classes. The formulation and derivation of this more general case is almost identical to that given above.

3 THE GROUPING EFFECT

A regression method is said to exhibit the "grouping effect" if the regression coefficients of a group of correlated features are approximately equal. In the analysis of the Elastic Net (Zou & Hastie, 2005), it was proven that for a given η and α, the absolute difference between any two identically-signed coefficients of the Elastic Net estimate $\hat{\beta}$ is correlation dependent. For sample correlation $\sigma_{ij} = X_i^T X_j$ it holds that:

$$|\hat{\beta}_i - \hat{\beta}_j| / \|y\|_2 \le (1/\eta\alpha)\sqrt{2(1-\sigma_{ij})}. \tag{14}$$

As σ_{ij} increases to 1, $|\hat{\beta}_i - \hat{\beta}_j| \to 0$. Similarly, we can bound the extent to which PEN groups variables. This is the content of next Theorem.

Theorem 3.1. *Let $P^\zeta = P - \zeta I$ and assume that P is* PSD *with nonnegative entries. Given that $sgn(\hat{\beta}_i) = sgn(\hat{\beta}_j)$, the Pairwise Elastic Net estimate $\hat{\beta}$ satisfies*

$$\frac{|\hat{\beta}_i - \hat{\beta}_j|}{\|y\|_2} \le \frac{1}{\sqrt{\eta\zeta^3}}\|P_i^\zeta - P_j^\zeta\|_2 + \frac{1}{\zeta\eta}\sqrt{2(1-\sigma_{ij})}$$

where $0 < \zeta \le \min_i P_{ii}$ and P_i^ζ denotes the i-th column of P^ζ.

Proof. Let $L(\beta) = (y - X\beta)^T(y - X\beta) + \eta|\beta|^T P|\beta|$. Since $\hat{\beta}$ is optimal, for $i = 1, \ldots, p$,

$$\tfrac{\partial L}{\partial \beta_i}|_{\hat{\beta}_i} = -2X_i{}^T(y - X\hat{\beta}) + 2\eta\,\mathrm{sgn}(\hat{\beta}_i)P^T|\hat{\beta}| = 0 \tag{15}$$

with P_i the i-th column of P. Consider components i and j of β with $\mathrm{sgn}(\hat{\beta}_i) = \mathrm{sgn}(\hat{\beta}_j)$. Subtracting (15) for i and j yields

$$2(X_j - X_i)^T(y - X\hat{\beta}) + 2\eta\ \mathrm{sgn}(\hat{\beta}_i)(P_i - P_j)^T|\hat{\beta}| = 0 \tag{16}$$

Now $P_i = \zeta e_i + P_i^\zeta$, $i = 1, \ldots, p$, where e_i is i-th standard basis vector. Hence $(P_i - P_j)^T|\hat{\beta}| = \zeta(|\hat{\beta}_i| - |\hat{\beta}_j|) + (P_i^\zeta - P_j^\zeta)^T|\hat{\beta}|$. Substituting this into (16) gives:

$$\hat{\beta}_i - \hat{\beta}_j = \frac{1}{\zeta}\mathrm{sgn}(\hat{\beta}_i)(P_j^\zeta - P_i^\zeta)|\hat{\beta}| + \frac{1}{\eta\zeta}(X_i - X_j)^T(y - X\hat{\beta}) \tag{17}$$

Taking the ℓ_2 norm of both sides of (17), applying the triangle inequality, and using $\|X_i - X_j\|_2^2 = 2(1 - \sigma_{ij})$ yields:

$$|\hat{\beta}_i - \hat{\beta}_j| \le \tfrac{1}{\zeta}\|P_j^\zeta - P_i^\zeta\|_2\|\hat{\beta}\|_2 + \tfrac{1}{\eta\zeta}\sqrt{2(1-\sigma_{ij})}\|y - X\hat{\beta}\|_2 \tag{18}$$

From the optimality of $\hat{\beta}$, $\|y\|_2^2 = L(0) \ge L(\hat{\beta}) = \|y - X\hat{\beta}\|_2^2 + \eta\zeta\|\hat{\beta}\|_2^2 + \eta|\hat{\beta}|^T(P - \zeta I)|\hat{\beta}|$. Since $P^\zeta = P - \zeta I$ contains nonnegative entries by construction, the previous expression implies that $\|y\|_2^2 \ge \|y - X\hat{\beta}\|_2^2$ and $\|y\|_2^2 \ge \eta\zeta\|\hat{\beta}\|_2^2$. Using these inequalities in (18) gives

$$|\hat{\beta}_i - \hat{\beta}_j| \le (\tfrac{1}{\sqrt{\eta\zeta^3}}\|P_j^\zeta - P_i^\zeta\|_2 + \tfrac{1}{\eta\zeta}\sqrt{2(1-\sigma_{ij})})\|y\|_2 \tag{19}$$

Dividing both sides by $\|y\|_2$ completes the proof. □

Both $1 - \sigma_{ij}$ and $\|P_i^\zeta - P_j^\zeta\|_2$ measure how similar features i and j are. The smaller these values are, the more similar features i and j. Thus the theorem states that the Pairwise Elastic Net assigns similar weights to highly correlated features, i.e., it exhibits the grouping effect. In the extreme case, for a correlation of 1 we have $P_i^\zeta = P_j^\zeta$, $\sigma_{ij} = 1$, and $\hat{\beta}_i = \hat{\beta}_j$.

At first glance we note that (14) is a tighter bound and so we can expect the Elastic Net to possess greater grouping capabilities. However, this is only one side of the argument. For $\sigma_{ij} \ll 1$ inequality (14) *also* bounds the difference of the coefficients. But for small σ_{ij} it is not desirable to have a tight bound for grouping. Thus PEN allows more slack in bounding this difference.

4 COORDINATE DESCENT

PEN can be recast as a quadratic program (QP) and solved using a QP solver. For moderately sized p, a QP solver is sufficient. As p increases, however, a different approach is needed. We propose a coordinate descent algorithm similar to those found in (Friedman et al., 2007). The approach is as follows: starting with the Lagrangian, $L(\beta)$, we evaluate $\partial L/\partial\beta_j$. Holding all β_j's fixed except for β_i we solve $\partial L/\partial\beta_i = 0$ for β_i and update accordingly. Cycling through each β_i iteratively will yield our solution.

Before proceeding we make note of the following equation for nonnegative scalars a and b:

$$az + b\,\mathrm{sgn}(z) = c \tag{20}$$

This equation is an indication of soft-thresholding, i.e.,

$$z = \frac{1}{a}\mathcal{S}(c, b) = \frac{1}{a}(|c| - b)_+\,\mathrm{sgn}(c) \tag{21}$$

In PEN, we have

$$L(\beta) = \|y - X\beta\|_2^2 + \eta|\beta|^T P|\beta| \tag{22}$$

$$= y^T y - 2q^T\beta + \beta^T Q\beta + \eta\sum_{i,j} P_{ij}|\beta_i\beta_j| \tag{23}$$

with P PSD and nonnegative, $Q = X^T X$, and $q = X^T y$. As exhibited before,

$$\frac{\partial L}{\partial \beta_i} = -2q_i + 2Q_i^T\beta + 2\eta\,\mathrm{sgn}(\beta_i)\sum_{j=1}^{p} P_{ij}|\beta_j| \tag{24}$$

Setting this partial derivative to 0, we solve for β_i in terms of $\beta_{-i} = \beta_{\{1:p\}\setminus i}$:

$$(Q_{ii} + P_{ii})\beta_i + \mathrm{sgn}(\beta_i)\eta\sum_{j\ne i} P_{ij}|\beta_j| = q_i - \sum_{j\ne i} Q_{ij}\beta_j \tag{25}$$

Algorithm 1 Coordinate Descent Algorithm for PEN

```
Input: $X$, $y$, $\eta$, $P$, $numiter$, $tol$, $\beta^0$
Initialize: $Q \leftarrow X^TX$, $q \leftarrow X^Ty$, $\beta^{old} \leftarrow \beta \leftarrow \beta^0$
for $j = 1$ to $numiter$ do
    for $i = 1$ to $p$ do
        $\beta_i \leftarrow \frac{\mathcal{S}\left(Q_{ii}\beta_i - Q_i^T\beta + q_i, \eta(P_i^T|\beta| - P_{ii}|\beta_i|)\right)}{Q_{ii}+\eta P_{ii}}$
    end for
    if $\|\beta - \beta^{old}\|_2 < tol$ then
        $\beta \leftarrow diag(1 + P_{ii}/Q_{ii})\beta$
        return
    end if
    $\beta^{old} \leftarrow \beta$
end for
```

which is of the form in (20). Thus we arrive at the update equation:

$$\beta_i \leftarrow \frac{\mathcal{S}\left(q_i - \sum_{j\neq i} Q_{ij}\beta_j, \eta\sum_{j\neq i} P_{ij}|\beta_j|\right)}{Q_{ii} + \eta P_{ii}} \tag{26}$$

which is equivalent to

$$\beta_i \leftarrow \frac{\mathcal{S}\left(Q_{ii}\beta_i - Q_i^T\beta + q_i, \eta(P_i^T|\beta| - P_{ii}|\beta_i|)\right)}{Q_{ii} + \eta P_{ii}} \tag{27}$$

The resulting algorithm is given by Algorithm 1.

5 RESCALING

In (Zou & Hastie, 2005) a distinction was made between the naive Elastic Net (NEN) solution and the Elastic Net solution. Whereas $\hat{\beta}_{nen}$ is the solution to (4), the EN solution is given by $\hat{\beta}_{en} = (1 + \lambda_2)\hat{\beta}_{nen}$. The reason for this rescaling is due to the "double shrinkage", as described by the authors. The first form of shrinkage stems from the ℓ_1 regularization, which results in soft-thresholding. The second form of shrinkage is contributed to the ℓ_2 regularization, which shrinks the correlation matrix toward identity. The soft-thresholding is desired because it yields sparse solutions. The tradeoff encountered in correlation shrinkage is that the problem is strictly convex and more than n features can be selected, but this is accompanied by the scaling of the coefficients. The $(1 + \lambda_2)$ factor corrects for this scaling.

The same "double shrinkage" occurs in PEN, but in a different way. The double shrinkage of NEN is made apparent by the coordinate descent update equation. Following the procedure in the previous section, we can show that the NEN update equation is given by

$$\beta_i \leftarrow \frac{\mathcal{S}\left(Q_{ii}\beta_i - Q_i^T\beta + q_i, \lambda_1/2\right)}{Q_{ii} + \lambda_2} \tag{28}$$

Comparing this to the update equation of OLS,

$$\beta_i \leftarrow \frac{Q_{ii}\beta_i - Q_i^T\beta + q_i}{Q_{ii}} \ , \tag{29}$$

we see that OLS normalization for β_i is simply Q_{ii}, i.e., the denominator. Therefore, once convergence has been reached, we can rescale the obtained NEN solution to match the same normalization as OLS. This involves multiplying by

$$\frac{Q_{ii} + \lambda_2}{Q_{ii}} = 1 + \frac{\lambda_2}{Q_{ii}} \ , \tag{30}$$

which is just $(1 + \lambda_2)$ when our input is standardized (recall: $Q_{ii} = X_i^TX_i = 1$). Applying the same reasoning for PEN we arrive at the rescaling for β_i:

$$\frac{Q_{ii} + \eta P_{ii}}{Q_{ii}} = 1 + \eta\frac{P_{ii}}{Q_{ii}} \ , \tag{31}$$

which is $(1 + \eta P_{ii})$ for standardized inputs.

6 EXPERIMENTS

In this section we discuss results from testing the PEN on a real data set of 97 measurements collected in a prostate cancer study by Stamey et al. (1989). This data set is a good benchmark for the Pairwise Elastic Net as it was previously examined by both Tibshirani (1996) and Zou & Hastie (2005). There are 8 predictors, log(cancer volume) (lcavol), log(prostate weight) (lpw), age, log(benign prostatic hyperplasia) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason), and percentage Gleason score 4 or 5 (pgg45). The response vector y is log(prostate specific antigen). As in Zou & Hastie (2005), the data were split into a training set of 67 and a test set of 30, and 10-fold cross-validation was used to set the tuning parameters. P was constructed as in (11), with $R_{ij} = |X_i^TX_j|$. The standard error was estimated from 1000 Bootstrap resamplings of the cross-validation error. The resulting mean-squared errors (MSE) are shown in Table 1. We see that the Pairwise Elastic Net achieves the same MSE as the Lasso and the Elastic Net do, within a standard error.

Table 1: Prostate Cancer Dataset Results

Method	**MSE(std. error)**
OLS	0.521 (0.116)
Ridge	0.489 (0.113)
Lasso	0.452 (0.117)
EN	0.452 (0.050)
PEN	0.464 (0.107)

Now we consider a simulated example which combines the grouping effect and the versatility of encoding

Table 2: Simulation Results

Meth.	**Acc. (se)**	**Spars. (se)**	**MSE (se)**
Ridge	8.35 (0.38)	239.95 (0.22)	1.16 (0.08)
Lasso	10.97 (1.90)	28.33 (3.95)	1.16 (0.08)
EN	8.44 (0.80)	38.15 (6.10)	1.12 (0.07)
PEN	8.10 (0.57)	29.37 (3.88)	1.11 (0.06)
GLasso	6.07 (0.09)	125.71 (26.22)	1.03 (0.05)

group structure in $J(\beta)$. We set $(n, p) = (100, 240)$. The rows of the design matrix are i.i.d. draws from the Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$. The correlation matrix $\Sigma \in \mathbb{R}^{p \times p}$ has entries $\Sigma_{ij} = \sigma^{|i-j|}$ with $\sigma = -0.99$. This encodes two antipodal clusters into Σ (the even-indexed and odd-indexed features). We set $\beta_{act} = [1, -1, 2, -2, 3, -3, 0, \ldots, 0]^T$ and form $y = X\beta_{act} + \epsilon$ where $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. With a strong grouping effect we would expect $\beta_i \approx (1 + 2 + 3)/120 = 0.05$ for i even and $\beta_i \approx -0.05$ for i odd. From an estimation standpoint the grouping effect might lead to minimal loss, but accuracy of the estimated β has been completely ignored. Even if we managed to group the non-zero β_i, the resulting estimated β is a smoothed estimate and accuracy cannot be guaranteed. For this case, we leverage the group structure and set P via equation (13). The tuning parameters were set with a training set of $n = 700$ and a validation set of $n = 300$, then using those parameters we ran 100 trials with $n = 100$. The results are shown in Table 2, where accuracy is measured by $\|\beta_{act} - \beta\|$. While Group Lasso does the best with regard to MSE and accuracy, it fails in the sparsity category because it selects whole groups, and here the groups are large. On the other hand PEN is about as sparse as Lasso and has better MSE and accuracy than the others. In fact, in 84% of the trials PEN was more accurate than EN. This indicates that the PEN grouping effect exhibits flexibility for in-cluster features as well as out-of-cluster features.

7 CONCLUSION

We have developed an approach of employing a user defined similarity measure to both group correlated variables and encourage sparsity. The Pairwise Elastic Net implements this approach with a convex regression problem that simultaneously groups similar features and promotes sparsity. The Pairwise Elastic Net encompasses the Elastic Net or mimics the Group Lasso behavior if a particular choice of similarity measure is made. Thus one can view the Pairwise Elastic Net as a generalization of the aforementioned methods, with an advantage that in the Pairwise Elastic Net, a data analyst has the freedom to select a similarity measure which best fits the problem of interest.

Since the penalty in the PEN is convex, the minimization problem can be solved with standard quadratic program solvers. Unfortunately, performance of standard solvers can become unsatisfactory as the dimensionality of the problem increases, so we have implemented a fast coordinate descent algorithm which solves the PEN efficiently.

Many interesting questions remain. For example, what similarity measure yields the best performance on a given data set. The Pairwise Elastic Net penalty introduces many more parameters than Elastic Net and while we provide suggestions for choosing the P matrix, work still needs to be done to find an effective method for doing so. In a semi-supervised regression scenario, the unlabeled measurements could be utilized to construct P. Another intriguing problem is developing an algorithm for the Pairwise Elastic Net that produces the entire regularization path, similar to LARS for the Lasso (Efron et al., 2004) or LARS-EN for the Elastic Net (Zou & Hastie, 2005).

References

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Efron, B., Hastie, T., Johnstone, L., & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, *32*, 407–499.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, *20*, 101–148.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, *1*(2), 302–332.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Lv, J., & Fan, J. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, *37*, 3498–3528.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients. *Journal of Urology*, *16*, 1076–1083.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol*, *58*(1), 267–288.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, *68*, 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, *67*, 301–320.