Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

Penalized Euclidean distance regression

Daniel Vasiliu[a], Tanujit Dey[b] and Ian L. Dryden[c]*

Received 3 November 2017; Accepted 3 December 2017

A method is introduced for variable selection and prediction in linear regression problems where the number of predictors can be much larger than the number of observations. The methodology involves minimizing a penalized Euclidean distance, where the penalty is the geometric mean of the ℓ_1 and ℓ_2 norms of regression coefficients. This particular formulation exhibits a grouping effect, which is useful for model selection in high-dimensional problems. Also, an important result is a model consistency theorem, which does not require an estimate of the noise standard deviation. An algorithm for estimation is described, which involves thresholding to obtain a sparse solution. Practical performances of variable selection and prediction are evaluated through simulation studies and the analysis of real datasets. © **2018 The Authors.** ***Stat*** **Published by John Wiley & Sons Ltd.**

Keywords: Euclidean distance; grouping; penalization; prediction; regularization; sparsity; variable screening

1 Introduction

High-dimensional regression problems are of great interest in a wide range of applications, for example, in analysing microarrays (Hastie et al., 2008; Fan et al., 2009), functional magnetic resonance images (Caballero Gaudes et al., 2013) and mass spectrometry data (Tibshirani et al., 2005). We consider the problem of predicting a single response Y from a set of p predictors $X_1, \ldots, X_p$, where p can be much larger than the number of observations n of each variable. If $p > n$, commonly used methods include regularization by adding a penalty to the least squares objective function or variable selection of the most important predictors.

A wide range of methods is available for achieving one or both of the essential goals in linear regression: accomplishing predictive accuracy and identifying pertinent predictive variables. There is a very large literature on high-dimensional regression methods, for example, introductions to the area are given by Hastie et al. (2008, Section 3.4) and James et al. (2013, Chapter 3). Earlier methods for high-dimensional regression include procedures that minimize a least squares objective function plus a penalty on the regression parameters. The methods include ridge regression (Hoerl & Kennard, 1970a, 1970b) with a squared ℓ_2 penalty, Lasso (Tibshirani, 1996) with an ℓ_1 penalty and the elastic net (Zou & Hastie, 2005) with a linear combination of ℓ_1 and squared ℓ_2 penalties. Alternative methods include the Dantzig selector (Candès & Tao, 2007), where the correlation between the residuals and predictors is bounded; sure independence screening (Fan & Lv, 2008), where predictors are initially screened using componentwise regression;

[a]**Department of Mathematics and Applied Mathematics, Virginia Commonwealth University, Richmond, Virginia, 23284, USA**
[b]**Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, 44195, USA**
[c]**School of Mathematical Sciences, The University of Nottingham, Nottingham, NG7 2RD, UK**
*__Email: ian.dryden@nottingham.ac.uk__

and square root Lasso (Belloni et al., 2011), which involves minimizing an empirical norm of the residuals with an ℓ_1 penalty.

In our method we use the Euclidean distance objective function plus a new norm penalty based on the geometric mean of the ℓ_1 and ℓ_2 norms of the regression parameters. The combination of Euclidean loss and geometric mean penalty is the main contribution of the paper. The advantage of our approach is that we are able to provide a pivotal recovery property, and in addition gain the grouping property of the Elastic Net (where regression coefficients of a group of highly correlated variables are very similar). The resulting penalized Euclidean distance (PED) method is shown to work well in a variety of settings. A particularly strong feature is that it works well when there are correlated designs with weak signal and strong noise.

2 Penalized Euclidean distance

We assume that the data are organized as an $n \times p$ design matrix X, and an n-dimensional response vector Y, where n is the number of observations and p is the number of variables. The columns of the matrix X are denoted by X_j, that is, $X_j = (x_{1j}, x_{2j} \ldots, x_{nj})^T$, $j = 1, \ldots, p$, and the regression parameters are $\beta = (\beta_1, \ldots, \beta_p)^T$. We assume that a vector of outcomes Y is modelled linearly as $Y = X\beta^* + \sigma\epsilon$, where β^* is the true parameter vector of dimension p, the expectation of ϵ is zero and its covariance matrix is the identity matrix. Thus, we assume that the expectation of the response $Y = (y_1, \ldots, y_n)^T$ depends only on a few variables, and so $X\beta^* = X^*\tilde{\beta}^*$, where the columns of the matrix X^* are a subset of the set of columns of the entire design matrix X, so X^* is associated with a subset of indices $\mathfrak{J}^* \subset \{1, 2, \ldots, p\}$ and $\tilde{\beta}^*$ is a subvector of β^* with the zero elements removed whose dimension is equal to the cardinality of $\mathfrak{J}^*$. In general, if we try to minimize $\|Y - X^*\tilde{\beta}^*\|$ over choices of $\mathfrak{J}^*$ and vectors $\tilde{\beta}^*$, the optimal choice of $\mathfrak{J}^*$ may not be unique because an under-determined system could have solutions with different sparsity patterns, even if the degree of the optimal sparsity (model size) is the same. However, in the signal reconstruction problem that we consider, where a penalty on the parameters is introduced, we will show that under some assumptions we can approximate β^* in probability. The cardinality of $\mathfrak{J}^*$ (denoted by $|\mathfrak{J}^*|$) is assumed to be less than the number of observations, and when p is greater than $|\mathfrak{J}^*|$, a real challenge is to detect the set of irrelevant columns, namely, the variables that correspond to the position of the null components of β^* and thus not needed for efficiently controlling the outcomes Y.

Our method involves minimizing the Euclidean distance (as a loss function, essentially equivalent to the empirical norm used in the square root Lasso) between Y and $X\beta$, with a penalty based on the geometric mean of the ℓ_1 and ℓ_2 norms. In particular, we minimize

$$L_{PED}(\lambda, \beta) = \|Y - X\beta\| + \lambda\sqrt{\|\beta\|\|\beta\|_1}, \tag{1}$$

where λ is the scalar regularization parameter, $\beta = (\beta_1, \beta_2 \ldots, \beta_p)^T$ is a vector in $\mathbb{R}^p$ (to be optimized over), $\|\beta\|^2 = \sum_{i=1}^p \beta_j^2$ is the squared ℓ_2 norm and $\|\beta\|_1 = \sum_{i=1}^p |\beta_j|$ is the ℓ_1 norm. The PED estimator $\hat{\beta}$ is defined as the minimizer of the objective function (1), that is, $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ and

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \{L_{PED}(\lambda, \beta)\}. \tag{2}$$

The penalty is proportional to the geometric mean of the ℓ_1 and ℓ_2 norms and has only one control parameter, λ.

An alternative, well-established method that involves a convex combination of ℓ_1 and ℓ_2^2 penalties is the elastic net (Zou & Hastie, 2005). The Lasso (Tibshirani, 1996) is a special case for this penalty, and so the elastic net combines the two methods. Our method also inherits important properties of Lasso and ridge regression but in a radically different way. The square root Lasso (Belloni et al., 2011) involves minimizing $L_{SQL}(\lambda, \beta) = \frac{1}{n}\|Y - X\beta\| + \frac{\lambda}{n}\|\beta\|_1$,

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

and so the first term for the PED estimator is the same as that of the square root Lasso multiplied by n, and it is just the penalty that differs. Belloni et al. (2011) have given a rationale for choosing the regularization parameter using a property called pivotal recovery, without requiring an estimate of the noise standard deviation.

The PED penalty is identical to the Lasso penalty for a single non-zero β_i, and so for very sparse models, behaviour like the square root Lasso is envisaged. We shall show that for our estimator there is a grouping effect for correlated variables, which is a property shared by the elastic net. A grouping effect occurs where highly correlated predictors X_j, X_k will give rise to very similar regression parameter estimates, that is, $\hat{\beta}_j \approx \hat{\beta}_k$.

With the application of a location transformation, both the design matrix X and the response vector Y can be centred; here, we say that a vector is centred if the mean of its values is zero. Considering a scaling transformation, each covariate X_j can be regarded as a point on the unit hypersphere S^{n-1} with a centring constraint. Throughout the paper, we assume that Y has been centred and the columns of X have been standardized as described above.

3 Theoretical results

3.1 General properties

The PED objective function enables variable selection under some mild compatibility conditions. The concept is based on the simple fact that the sum of the squares of the relative sizes of vector components (as defined by $\beta_j/\|\beta\|$) is always equal to 1. For any vector in $\mathbb{R}^p$, if there are components that have relative sizes larger than $\frac{1}{\sqrt{p}}$, then the other components must have relative sizes falling under this value. In addition, if many components have similar relative sizes due to a grouping effect, then the relative size of those components must be small. The new penalty function that we consider is actually a norm.

Lemma 1

The geometric mean of ℓ_1 and ℓ_2 norms is also a norm on finite-dimensional Euclidean vector spaces.

The following theorem demonstrates the grouping effect achieved by a minimizer of the PED objective function. The idea of grouping effect was first introduced by Zou & Hastie (2005). Our version of the grouping effect involves the relative contributions of the components of the minimizer of the PED objective function. This property enables the process of eliminating irrelevant variables from the model, and considering the situation of $p > n$, the process of selecting and grouping variables is an important priority. Theorem 1 supports the idea of obtaining groups of highly correlated variables, based on the relative size of the corresponding component minimizers of the PED objective function.

Theorem 1

Assume we have a standardized data matrix X and Y is a centred response vector. Let $\hat{\beta}$ be the PED estimate given by $\hat{\beta}(\lambda) = \arg\min_{\beta}\{L_{PED}(\lambda,\beta)\}$ for some $\lambda > 0$. Define $D_\lambda(i,j) = \frac{1}{\|\hat{\beta}(\lambda)\|}|\hat{\beta}_i(\lambda) - \hat{\beta}_j(\lambda)|$; then

$$D_\lambda(i,j) \leq \frac{2(1-\rho_{ij})^{1/2}}{\lambda} \leq \frac{2\theta_{ij}}{\lambda},$$

where $\rho_{ij} = (X_i)^T(X_j)$ is the sample correlation and θ_{ij} is the angle between X_i and X_j, $0 \leq \theta_{ij} \leq \pi/2$.

Note that this result is analogous to Theorem 1 of Zou & Hastie (2005) for the elastic net, and the same method of proof is used in the Appendix. From Theorem 1, if θ_{ij} is small, then the corresponding parameters estimated from PED regression will be similar, which is the grouping effect. When $\theta_{ij} = 0$, we have the following corollary.

Corollary 1
Let $\hat{\beta}(\lambda) = \arg\min_{\beta}\{L_{PED}(\lambda, \beta)\}$. *If* $X_i = X_j$, *then* $\hat{\beta}_i(\lambda) = \hat{\beta}_j(\lambda)$.

The grouping effect means that if we have strong overcrowding on the unit hypersphere around an irrelevant column, then this would be detected by a large drop in the relative size of the corresponding components of the solution to our objective function.

We consider the case when the number of variables by far exceeds the number of true covariates. Therefore, the cardinality of the set $\mathcal{S}$ is infinite, and the challenge is to find a sparse solution in it. The starting point of our analysis will be a solution of the PED problem defined by (2). As before, we let $\hat{\theta}_j$ represent the angle between vectors X_j and $Y - X\hat{\beta}$. We note that the angle $\hat{\theta}_j$ satisfies the equation $\hat{\theta}_j = \frac{\pi}{2} - \arcsin\left(\frac{X_j^T(Y - X\hat{\beta})}{\|Y - X\hat{\beta}\|}\right)$, $0 \leq \hat{\theta}_j < \pi$, whenever $\|Y - X\hat{\beta}\| \neq 0$. Define $\hat{k} := \left(\frac{\|\hat{\beta}\|}{\|\hat{\beta}\|_1}\right)^{1/2}$; we have $\frac{1}{\sqrt[4]{p}} \leq \hat{k} \leq 1$ provided $\hat{\beta} \neq 0_{\mathbb{R}^p}$. Note that $\hat{k}$ is a measure of sparsity, with the highest value 1 when there is a single non-zero element in β (very sparse) and with the smallest value when all elements of β are equal and non-zero (very non-sparse). We assume that $0_{\mathbb{R}^p}$ is not a minimizer of $\|Y - X\beta\|$.

Lemma 2
If $\hat{\beta}(\lambda)$ *is a solution of* (2) *and* $\hat{\beta}_j(\lambda) \neq 0$, *then* $\frac{\hat{\beta}_j(\lambda)}{\|\hat{\beta}(\lambda)\|} = \hat{k}\left(\frac{2\cos(\hat{\theta}_j)}{\lambda} - \hat{k}\ \operatorname{sgn}(\hat{\beta}_j(\lambda))\right)$.

Result 1
We have $|\cos(\hat{\theta}_j)| \leq \frac{\lambda\hat{k}}{2}$ *if and only if* $\hat{\beta}_j(\lambda) = 0$.

Result 2
If $\hat{\beta}$ *is the solution of* (2) *and its jth component is non-zero (i.e.* $\hat{\beta}_j \neq 0$*), then* $\operatorname{sgn}(\hat{\beta}_j) = \operatorname{sgn}(X_j^T(Y - X\hat{\beta})) = \operatorname{sgn}(\frac{\pi}{2} - \hat{\theta}_j)$.

The following result helps demonstrate the existence of a minimizing sequence whose terms have the grouping effect property for the relative size of their components.

Lemma 3
If $\hat{\beta}$ *is the solution of* (2), *then* $\left|\frac{\hat{\beta}_j}{\|\hat{\beta}(\lambda)\|}\right| < M \leq 1$ *if and only if* $|\cos(\hat{\theta}_j)| \leq \frac{\lambda}{2}\left(\hat{k} + \frac{M}{\hat{k}}\right)$, *where M is a constant.*

3.2 Model consistency

In this section, we demonstrate that our method is also able to recover sparse signals without (pre-)estimates of the noise standard deviation or any knowledge about the signal. In Belloni et al. (2011), this property is referred to as pivotal recovery. An important aspect is that an oracle theorem also brings a solid theoretical justification for the choice of the parameter λ.

We assume that $Y = X\beta^* + \sigma\epsilon$, where β^* is the unknown true parameter value for β, σ is the standard deviation of the noise and ϵ_i, $i = 1, \ldots, n$, is independent and identically distributed with a normal law Φ_0 with $E_{\Phi_0}(\epsilon_i) = 0$ and $E_{\Phi_0}(\epsilon_i^2) = 1$. Let $\mathfrak{J}^* = \operatorname{supp}(\beta^*)$. For any candidate solution $\hat{\beta}$, we can use the notation L for the plain Euclidean distance $L(\hat{\beta}) = \|Y - X\hat{\beta}\|$, and the newly introduced norm is denoted by $\|\beta\|_{(1,2)}$, that is, $\|\beta\|_{(1,2)} = (\|\beta\|_1\|\beta\|)^{1/2}$.

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

The idea behind the following considerations is the possibility of estimating the quotient $\|X^T\epsilon\|_\infty/\|\epsilon\|$ in probability following Belloni et al. (2011). We can use the same general result to show that the method we propose is also capable of producing pivotal recoveries of sparse signals.

Before stating the main theorem, we introduce some more notations and definitions. The solution of the PED objective function is denoted by $\hat{\beta}(\lambda)$. Let $\|u\|_X$ denote $\|Xu\|$, p^* the cardinality of $\mathfrak{J}^*$, $M^* = \|\beta^*\|$, $S = \|X^T\epsilon\|_\infty/\|\epsilon\|$, $c > 1$ and, for brevity, $\bar{c} = (c+1)/(c-1)$. Also, we write u^* for the vector of components of u that correspond to the non-zero β^* elements, that is, with indices in $\mathfrak{J}^*$. Also, we write u^{*c} for the vector of components of u that correspond to the zero elements of β^*, that is, with indices in the complement of $\mathfrak{J}^*$. We shall initially focus on the case $n^2 > p$. Consider

$$\Delta_{\bar{c}}^* = \left\{u \in \mathbb{R}^p : u \neq \{0_{\mathbb{R}^p}\},\ \|u^{*c}\|_1 \leq \bar{c}\|u^*\|_1 + \frac{c\sqrt[4]{p}}{c-1}\sqrt[4]{p^*}M^*\right\}. \tag{3}$$

Assume that $\bar{k}_{\bar{c}}^* := \min_{u\in\Delta_{\bar{c}}^*} \frac{1}{\sqrt{n}}\frac{\|u\|_X}{\|u\|}$ and $k_{\bar{c}}^* := \left(1-\frac{1}{c}\right)\min_{u\in\Delta_{\bar{c}}^*} \frac{\sqrt{p^*}\|u\|_X}{2\|u^*\|_1 + \sqrt[4]{p}\sqrt[4]{p^*}M^*}$ are bounded away from 0. We make the remark that if the first compatibility condition holds, there is a relatively simple scenario when the second condition would hold as well. If $\bar{k}_{\bar{c}}^*$ is bounded away from 0 on $\Delta_{\bar{c}}^*$, we have that $\|u\|_X$ must be at least $O(\sqrt{n})$ on $\Delta_{\bar{c}}^*$. At the same time, if $\|u^*\|$ is at most $O(p^*)$, we therefore have

$$k_{\bar{c}}^* = \sqrt{p^*}O\left(\sqrt{n}\right)/\left(O(p^*) + \sqrt[4]{p}\sqrt[4]{p^*}M^*\right) = O\left(\sqrt{n}\right)/\left(O\left(\sqrt{p^*}\right) + \sqrt[4]{p}\sqrt[4]{p^*}\left(M^*/\sqrt{p^*}\right)\right),$$

and we assume $M^*/\sqrt{p^*}$ is bounded. Thus, the second compatibility condition could be easily achieved in the case when $p = n^{1+\alpha_1}$ and $p^* = n^{\alpha_2}$ with $\alpha_1, \alpha_2 > 0$ and $\alpha_1 + \alpha_2 \leq 1$. We also present a result with certain compatibility conditions for the case when $p > n^2$ in Vasiliu et al. (2017).

We refer to $k_{\bar{c}}^*$ and $\bar{k}_{\bar{c}}^*$ as restricted eigenvalues. The concept of restricted eigenvalues was introduced by Bickel et al. (2009) with respect to the ℓ_1 penalty function. Our definition and usage are adapted to our own objective function. As stated before, our oracle theorem is based on the estimation of $\frac{\|X^T\epsilon\|_\infty}{\|\epsilon\|}$. Directly following from Lemma 1 of Belloni et al. (2011), we have the following lemma.

Lemma 4
Given $0 < \alpha < 1$ and $c > 1$, the choice $\lambda = \frac{c\sqrt[4]{p}}{\sqrt{n}}\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)$ satisfies $\lambda \geq c\sqrt[4]{p}S$ with probability $1-\alpha$.

Now we are ready to state the main result.

Theorem 2
Assume that $\lambda \leq \frac{\rho\sqrt[4]{p}k_{\bar{c}}^*}{\sqrt{p^*}}$ for some $0 < \rho < 1$. If also $\lambda \geq c\sqrt[4]{p}S$, then

$$(1-\rho^2)\|u\|_X \leq \frac{2c\sqrt{p^*\log(2p/\alpha)}L(\beta^*)}{k_{\bar{c}}^*\sqrt{n}}. \tag{4}$$

If $L(\beta^*) = O_p(\sqrt{n})$ and $\sqrt{p^*\log(2p/\alpha)}/\sqrt{n} \to 0$, a direct consequence is that $\hat{\beta}(\lambda) \to \beta^*$ in probability.

We can use the value of λ in Lemma 4 for practical implementation in order to ensure $\lambda \geq c\sqrt[4]{p}S$ holds with probability $1-\alpha$. Note that the rate of convergence is asymptotically the same as rates seen in other sparse regression problems (e.g. Negahban et al., 2012), although as for the square root Lasso of Belloni et al. (2011), knowledge of σ is not needed. Also, there are some circumstances when we can consider other values of λ.

Corollary 2

Let $0 < \xi < 1$ *and* $\Delta_\xi = \left\{u \in \mathbb{R}^p, \frac{\sqrt{n}}{\sqrt[4]{p}}\|u^{*c}\|_1 \xi \leq \|u^*\|_1 \left(\frac{2\sqrt{n}}{\sqrt[4]{p}} - \xi\right) + \|\beta^*\|_1 \left(1 - \frac{\sqrt{n}}{\sqrt[4]{p}}\right)\right\}$. *Assuming that* $k^*_\xi := \min_{u \in \Delta_\xi} \frac{\frac{\sqrt{p^*}}{\sqrt{n}}\|u\|_X}{\frac{2\sqrt{n}}{\xi\sqrt[4]{p}}\|u^*\|_1 + \frac{\|\beta^*\|_1}{\xi}\left(1 - \frac{\sqrt{n}}{\sqrt[4]{p}}\right)} > k > 0$, *and for* $\lambda = c\Phi_0^{-1}\left(1 - \frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{n}$, *with* $c > 1$, *if* $\sqrt{\frac{\|\hat{\beta}(\lambda)\|}{\|\hat{\beta}(\lambda)\|_1}} - \frac{\sqrt{n}}{c\sqrt[4]{p}} \geq \xi > 0$ *and, at the same time, we assume* $\lambda \leq \frac{\rho\sqrt[4]{p}k^*_\xi}{\sqrt{n}\sqrt{p^*}}$ *for some* $0 < \rho < 1$, *then we also have an oracle property, that is,*

$$\|\hat{\beta}(\lambda) - \beta^*\| = O\left(\sqrt{\frac{p^* \log(2p/\alpha)}{n}}\right)$$

with probability $1 - \alpha$.

We use the corollary to suggest a method for choosing the model parameters by maximizing $\hat{k} = \left(\frac{\|\hat{\beta}(\lambda)\|}{\|\hat{\beta}(\lambda)\|_1}\right)^{1/2}$, and this would encourage the selection of sparse models.

For a practical implementation of our method, we make use of the proven theoretical results. From the signal recovery theorem and corollary, we obtain that $\|\hat{\beta}(\lambda) - \beta^*\| = O(\sqrt{p^* \log(2p/\alpha)}/\sqrt{n})$ with probability $1-\alpha$. Thus, if j is an index where there is no signal, that is, $\beta^*_j = 0$, then, from the previous equation, we have that $|\hat{\beta}_j(\lambda))| < \|\hat{\beta}(\lambda) - \beta^*\| \leq$ const. $\sqrt{p^* \log(2p/\alpha)}/\sqrt{n}$. If $\|\hat{\beta}(\lambda)\| \neq 0$, we can divide by $\|\hat{\beta}(\lambda)\|$ and obtain

$$\frac{|\hat{\beta}_j(\lambda))|}{\|\hat{\beta}(\lambda))\|} < \delta(p)/\sqrt{n}, \text{where } \delta(p) \propto \sqrt{p^* \log(2p/\alpha)}. \tag{5}$$

We will use (5) to inform a threshold choice as part of the PED numerical implementation in the next section. As well as dependence on n, we also investigate the effect of p on the relative size of the components.

3.3 Theoretical comparisons

In Section 3.1, we presented general properties for the PED estimator without requiring any special assumptions. The elastic net also shares the grouping effect without any special assumptions, but the Lasso and square root Lasso do not. The other results in Section 3.1 are particular to our estimator. In Section 3.2, under special assumptions, the Lasso, square root Lasso, and PED estimators have near-oracle rates of the same order, which is $O(\sqrt{p^* \log(p)/n})$. The main difference is that in the case of the Lasso pre-estimates of the noise standard deviation, σ is needed as well as different restricted eigenvalue assumptions given by Bickel et al. (2009). For the square root Lasso and PED, the near-oracle rate can be achieved without the pre-estimates of the noise, but with PED, we also do guarantee the grouping effect that is mitigating multiple correlations. The simulations below show numerical evidence that our rate could actually be better than the one used as a theoretical benchmark for comparison with Lasso and square root Lasso.

4 Numerical implementation

The objective function $L_{PED}(\lambda, \beta) = \|Y - X\beta\| + \lambda(\|\beta\|\|\beta\|_1)^{1/2}$ is convex for any choice of λ and also differentiable on all open orthants in $\mathbb{R}^p$ bounded away from the hyperplane $Y - X\beta = 0$. In order to find good approximations for minimizers of our objective function, as in many cases of non-linear large-scale convex optimization problems, a quasi-Newton method may be preferred because it is known to be considerably faster than methods like coordinate descent

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

by achieving super-linear convergence rates. Another important advantage is that second derivatives are not necessarily required. For testing purposes, we present a numerical implementation based on the well-performing quasi-Newton methods for convex optimization known as Broyden–Fletcher–Goldfarb–Shanno (BFGS) methods: limited-memory BFGS (L-BFGS) (Nocedal, 1980) and BFGS (Bonnans et al., 2006). We also tested a version of non-smooth BFGS called the hybrid algorithm for non-smooth optimization (Lewis & Overton, 2008) and obtained very similar results.

The idea for the estimation is to use theoretically informed parameters based on Theorem 2 and Corollaries 2 and (5), in order to choose a suitable value of λ and give a sparse estimate of β^* after thresholding. We are choosing a lambda value in the interval between $\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{n}$ and $\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{\sqrt{n}}$ that is maximizing $\hat{k}(\lambda_0) := \left(\frac{\|\hat{\beta}(\lambda_0)\|}{\|\hat{\beta}(\lambda_0)\|_1}\right)^{1/2}$. We retain the components of the solution that have higher relative contributions, that is, $\frac{|\hat{\beta}_j|}{\|\hat{\beta}\|} \geq \delta(p)/\sqrt{n}$, where $\delta(p)$ is a tuning thresholding constant that could be selected by some information criterion such as the Akaike information criterion (AIC) or by n-fold cross-validations, or we could fix δ, for example, $\delta = 0.75$. The steps for the numerical approximation of β^* by using the PED method are as follows:

1. Use a quasi-Newton algorithm (e.g. L-BFGS) to minimize the convex objective function (1) with λ values between $\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{n}$ and $\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{\sqrt{n}}$ and evaluate $\hat{k}(\lambda)$.
2. For the solution $\hat{\beta}$ that maximizes $\hat{k}(\lambda)$, set $\hat{\beta}_j = 0$ if $\frac{|\hat{\beta}_j|}{\|\hat{\beta}\|} \leq \delta(p)/\sqrt{n}$ (the choice of $\delta(p)$ is motivated by (5)). Eliminate the columns of the design matrix corresponding to the zero coefficients $\hat{\beta}_j$, with p^* columns remaining.
3. Use the quasi-Newton algorithm to minimize the objective function with the remaining columns of the design matrix and λ between $\Phi_0^{-1}\left(1-\frac{\alpha}{2p}\right)\frac{\sqrt[4]{p}}{n}$ and $\frac{\sqrt[4]{p^*}}{\sqrt{n}}$ and output the solution.

For all the numerical simulations and almost all real datasets, a default value of λ was used for the last step of the numerical approximation, namely, $\frac{\sqrt[4]{p^*}}{\sqrt{n}}$.

5 Numerical applications

5.1 Simulation study

We consider a simulation study to illustrate the performance of our method and the grouping effect in the case when $p \geq n$. In this example, we compare the results with the square root of the Lasso method (Belloni et al., 2011) that uses a scaled Euclidean distance as a loss function plus an ℓ_1 penalty term, using the asymptotic choice of λ. We also compare the results with both Lasso and elastic net methods as they are implemented in the publicly available packages for R, again using the default options. In particular, we used 10-fold cross-validation to choose the roughness penalty for the Lasso and elastic net using the command `cv.glmnet` in the R package `glmnet`, and we use the command `slim` in the R package `flare` with penalty term $\lambda = 1.1\Phi_0^{-1}(1-0.05/(2p))/\sqrt{n}$. We use the PED method with a default $\delta = 0.75$ or chosen with the AIC from a range of values between 0.75 and 1.5.

We consider situations with weak signal, strong noise and various correlated designs. In particular, for a range of values of n, p, and ρ, the data are generated from the linear model $Y = X\beta^* + \sigma\epsilon$, where

$$\beta^* = (\underbrace{0.3,\ldots,0.3}_{4},\underbrace{0,\ldots,0}_{50},\underbrace{0.3,\ldots,0.3}_{4},\underbrace{0,\ldots,0}_{50},\underbrace{0.3,\ldots,0.3}_{4},0,\ldots,0)^T,$$

$\|\beta^*\|_0 = 12$, $\sigma = 1.5$ and X generated from a p-dimensional multivariate normal distribution with mean zero and correlation matrix Σ, where the (j,k)th entry of Σ is $\rho^{|j-k|}$, $1 \leq j,k \leq p$.

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

Stat
The ISI's Journal for the Rapid Dissemination of Statistics Research

The results are summarized in Tables I and II, and the reported values are based on averaging over 100 datasets. The distance between the true signal and the solution produced is also recorded using root mean square error. Under highly correlated designs, the method we propose shows a very efficient performance against the "curse of dimensionality" and overcrowding.

PED (as indicated in Table I) has performed very well, obtaining the highest rate of true positives in many examples. Also, we compare it with PED when the parameters are selected by AIC, given by PED(AIC) in Table I, which also performs well. The elastic net is the next best performance, and Lasso (for the strongly correlated case) and square

Table I. Simulation results based on the example in Section 5.1.

	$\rho = 0.5$			$\rho = 0.9$			$\rho = 0.99$		
	TP	MS	RMSE	TP	MS	RMSE	TP	MS	RMSE
$n = 100, p = 200$									
PED	**9.96**	20.80	**0.775**	**11.62**	26.97	**0.664**	**10.17**	39.28	**0.862**
PED(AIC)	9.37	17.37	0.789	10.57	17.87	0.692	7.07	15.87	1.013
Elastic net	9.34	30.17	0.949	9.43	25.63	1.064	6.2	23.52	1.436
Lasso	8.75	26.18	0.879	7.38	20.15	1.069	3.59	14.34	1.637
Square root Lasso	1.18	1.2	1.015	3.94	4.65	0.927	3.03	8.84	1.291
$n = 100, p = 1{,}000$									
PED	**9.53**	47.34	1.049	**11.52**	35.44	**0.706**	**11.14**	72.19	**0.847**
PED(AIC)	9.34	43.72	1.051	10.48	23.93	0.741	8.87	58.94	1.012
Elastic net	7.93	44.48	1.042	9.46	37.73	1.077	7.51	34.46	1.265
Lasso	7.07	32.27	**0.956**	6.87	27	1.044	3.3	18.03	1.510
Square root Lasso	0.9	0.94	1.025	2.91	3.42	0.965	2.97	9.07	1.253
$n = 200, p = 200$									
PED	**11.71**	21.19	**0.581**	**11.97**	27.15	0.540	**11.46**	44.47	0.781
PED(AIC)	11.62	20.31	0.591	11.83	20.3	**0.535**	10.27	25.09	**0.777**
Elastic net	11.41	34.3	0.687	10.45	25.57	0.860	7.72	23.53	1.117
Lasso	11.03	30.89	0.653	8.85	21.59	0.900	5.04	15.3	1.396
Square root Lasso	5.42	5.47	0.900	7.26	8.3	0.830	4.84	11.32	1.190
$n = 200, p = 2{,}000$									
PED	**11.21**	65.23	0.880	**11.76**	31.79	**0.556**	**11.99**	117.42	**0.838**
PED(AIC)	11.20	61.35	0.877	11.36	23.95	0.577	9.52	52.21	0.908
Elastic net	10.52	57.22	0.794	10.79	42.47	0.835	9.55	37.03	1.007
Lasso	9.79	44.42	**0.752**	8.65	31.19	0.856	4.77	19.74	1.280
Square root Lasso	3.57	3.57	0.961	6.45	7.09	0.837	4.8	11.14	1.155
$n = 200, p = 3{,}000$									
PED	**11.17**	73.91	0.977	**11.79**	35.04	**0.581**	**11.7**	121.92	**0.854**
PED(AIC)	11.16	72.10	0.975	11.66	27.69	0.586	9.39	77.3.02	0.934
Elastic net	10.23	64.83	0.834	10.75	44.16	0.839	9.94	37.42	0.993
Lasso	9.61	49.97	**0.788**	8.48	33.77	0.870	4.62	19.04	1.293
Square root Lasso	2.86	2.87	0.980	6.31	7.05	0.843	4.6	10.79	1.173

Note: The TPs are the average number of non-zero parameters, which are estimated as non-zero and the MS is the average number of estimated non-zero parameters, from 100 simulations. The RMSE is given for estimating β^*. The best values in the TP and RMSE columns are in bold. AIC, Akaike information criterion; MS, model size; PED, penalized Euclidean distance; RMSE, root mean square error; TP, true positive.

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

Table II. Simulation results based on example 1 and the numerical implementation described in the previous section.

	$\rho = 0.5$			$\rho = 0.9$			$\rho = 0.99$		
	$MS = 10$	$MS = 20$	$MS = 30$	$MS = 10$	$MS = 20$	$MS = 30$	$MS = 10$	$MS = 20$	$MS = 30$
$n = 100, p = 200$									
PED	**7.10**	**9.86**	**11.21**	**6.79**	**9.81**	**12**	**4.98**	**7.33**	**8.26**
Elastic net	6.54	8.55	9.53	6.51	8.27	9.04	3.02	5	6.4
Lasso	6.13	8.14	9.16	5.39	6.75	7.39	2.81	3.42	4.05
Square root Lasso	6.17	8.18	9.04	6.69	7.52	7.58	3.33	4.02	4.05
$n = 100, p = 1{,}000$									
PED	**6.88**	**8.94**	**10.24**	**6.84**	**9.39**	**11.34**	**4.61**	**7.18**	**8.15**
Elastic net	5.37	6.98	7.84	5.37	6.18	6.93	3.18	5.12	6.51
Lasso	5.11	6.56	7.20	5.42	6.27	6.66	2.85	3.43	3.75
Square root Lasso	5.75	7.01	7.43	5.37	6.18	6.76	3.39	3.67	3.9
$n = 200, p = 200$									
PED	**8.82**	**11.11**	**11.59**	7.51	**10.27**	**11.95**	**6.57**	**9.62**	**11.17**
Elastic net	8.19	9.95	10.39	**7.82**	9.59	10.19	4.88	7.09	8.5
Lasso	7.64	9.21	9.62	7.14	8.27	8.93	3.5	4.37	4.96
Square root Lasso	8.53	10.55	11.11	7.00	8.32	8.85	3.30	4.41	5.19
$n = 200, p = 2{,}000$									
PED	**7.92**	**9.60**	**10.04**	7.47	**10.19**	**12**	**6.08**	**8.76**	**9.74**
Elastic net	7.03	8.34	8.95	**7.85**	9.48	9.94	4.08	5.98	7.41
Lasso	6.96	8.45	8.99	6.78	7.78	8.15	3.91	4.29	4.57
Square root Lasso	7.47	9.01	9.72	7.02	8.25	8.77	3.89	4.52	4.72
$n = 200, p = 3{,}000$									
PED	**8.99**	**11.35**	**11.95**	7.2	**10.5**	**12**	**6.77**	**9.24**	**10.07**
Elastic net	7.36	8.6	9.09	**8.13**	9.89	10.21	4.46	6.65	8.14
Lasso	7.13	8.32	8.66	6.49	7.31	7.61	4.09	4.61	4.84
Square root Lasso	7.35	8.86	9.4	7.11	8.15	8.51	4.08	4.65	4.89

Note: The best values for the true positives are in bold. The true positives are the average number of non-zero parameters that are estimated as non-zero and the fixed MS from 100 simulations. MS, model size; PED, penalized Euclidean distance.

root Lasso have low rates of true positives. Note that the square root Lasso has performed rather differently here from the others. It is the only method using an asymptotic value of λ, where n may not be large enough here. PED has a lower model size compared to the elastic net. Finally, the root mean square error is generally best for PED, particularly for the more highly correlated situations. Table II summarizes the true positive rates for fixed model sizes, where the largest k_{MS} coefficients in absolute value are retained for model size k_{MS}. Again, PED has performed very well in retaining the highest number of true positives in nearly all cases (42 out of 45), with elastic net being the best in three cases. Overall, PED has performed extremely well in these simulations.

5.2 Real data applications

For prediction performance comparison, we considered the datasets *Air* (Chambers et al., 1983), *Servo* (Quinlan, 1993; Lichman, 2013), *Tecator* (Borggaard & Thodberg, 1992), *Housing* (Harrison & Rubinfeld, 1978), *Ozone* (Breiman & Friedman, 1985) and *Iowa* (Draper & Smith, 1998). It is important to note that throughout the simulations and the real data analyses, both Lasso and elastic net were run with double cross-validation for selecting the model

Table III. Prediction mean square error rates using 10-fold cross-validation.

	PED	PED-CV	Elastic net	Lasso	Square root Lasso
Servo	0.3436	**0.2977**	0.3041	0.3130	0.7393
Pollute	2,334.0	**2,197.3**	2,334.8	2,778.6	3,066.9
Iowa	**89.3389**	100.8377	143.8566	122.8793	133.9454
Air	**0.2348**	0.2413	0.2811	0.2874	0.3962
Ozone	17.8964	**15.6248**	16.8717	17.1563	21.4246
Tecator	12.0489	**10.8289**	50.0329	44.3648	131.1065

Note: The smallest values are in bold. CV, cross-validation; PED, penalized Euclidean distance.

size and the tuning parameter. We used PED with default values, and we also ran PED with a single 10-fold cross-validation for tuning δ that affects only the model size; the results are reported in Table III. In some cases, such as the Tecator dataset, the prediction error further improved when the final choice of λ was $\Phi_0^{-1}(1-\frac{\alpha}{2p})\frac{\sqrt[4]{p}}{n}$. Also, in the case of the Servo dataset, the variable selection benefitted from searching for lambda in a subinterval of the one proposed by default, namely, between $\Phi_0^{-1}(1-\frac{\alpha}{2p})\frac{\sqrt[4]{p}}{n}$ and $\frac{\sqrt[4]{p}}{\sqrt{n}}$. PED has again performed well, particularly when using cross-validation.

Melanoma: In this application, we implement PED as a variable selection tool when the response variable serves as binary classification. We consider an application of the method to a proteomics dataset from the study of melanoma (skin cancer). The mass spectrometry dataset was described by Mian et al. (2005) and further analysed by Browne et al. (2010). The data consist of mass spectrometry scans from serum samples of 205 patients, with 101 patients with stage I melanoma (least severe) and 104 patients with stage IV melanoma (most severe). Each mass spectrometry scan consists of an intensity for 13,951 mass over charge (m/z) values between 2,000 and 30,000 Da. It is of interest to find which m/z values could be associated with the stage of the disease, which could point to potential proteins for use as biomarkers. We first fit a set of 500 important peaks to the overall mean of the scans using the deterministic peak-finding algorithm of Browne et al. (2010) to obtain 500 m/z values at peak locations. We consider the disease stage to be the response, with $Y=-1$ for stage I and $Y=1$ for stage IV. Note that we have an ordered response here as stage IV is much more severe than stage I, and it is reasonable to treat the problem as a regression problem.

We fit the PED regression model versus the intensities at the 500 peak locations. We have $n=205$ by $p=500$. The data are available at http://www.maths.nottingham.ac.uk/~ild/mass-spec

Here, we use $\alpha=0.05$. The parameter values chosen to maximize $\hat{k}$ are $\lambda=0.5$ and $\delta(p)=0.75$, selecting 96 non-zero m/z values. Browne et al. (2010) also considered a mixed-effects Gaussian mixture model and a two-stage t-test for detecting significant peaks. If we restrict ourselves to the coefficients corresponding to the 50 largest peaks in height, Browne et al. (2010) identified 17 as non-zero as did PED, with eight out of the 17 in common. If we apply PED(AIC), then seven peaks are chosen out of the largest 50, of which only two are in common with Browne et al. (2010). The elastic net chose six peaks with five of those in common with Browne et al. (2010), and for the Lasso, five peaks were chosen from the top 50 largest, with four in common with Browne et al. (2010). Note that here PED has selected the most peaks in common with Browne et al. (2010), and it is reassuring that the different methods have selected some common peaks.

6 Appendix: Proofs

Proof

Lemma 1 □

For $\beta \in \mathbb{R}^p$, let $\ell_1(\beta)$ denote $\|\beta\|_1$ and $\ell_2(\beta)$ denote $\|\beta\|$. Consider the set

$$\mathcal{C} = \{\beta \in \mathbb{R}^p : \sqrt{\ell_1(\beta)\ell_2(\beta)} \leq 1\}.$$

We notice that $\mathcal{C} \equiv \{\beta \in \mathbb{R}^p : \ell_1(\beta)\ell_2(\beta) \leq 1\}$, and therefore, $\mathcal{C}$ is a bounded, closed and convex subset of $\mathbb{R}^p$, which contains the origin. Let $g(\beta) = \ell_1(\beta)\ell_2(\beta)$, and let Epi$(g)$ *denote its epigraph, that is,* Epi$(g) = \{(\beta, t) \in \mathbb{R}^{p+1} : g(\beta) \leq t\}$. *The set $\mathcal{C}$ is convex and orthant symmetric. Indeed, the Hessian of g is positive semi-definite on each orthant of $\mathbb{R}^n$ because, after differentiating g twice with the product rule, it can be written as a sum of three matrices, which can be argued, by applying Sylvester's theorem, to be positive semi-definite. We see that in our case* Epi$(g) = \{t(\mathcal{C}, 1) : t \in [0, +\infty)\}$, *and therefore,* Epi$(g)$ *is a convex cone in $\mathbb{R}^{p+1}$ because $\mathcal{C}$ is a convex set in $\mathbb{R}^p$. This shows that $\sqrt{\ell_1\ell_2}$ is a convex function. Because $\sqrt{\ell_1\ell_2}$ is convex and homogeneous of degree 1, it follows that it must also satisfy the triangle inequality. Therefore, $\sqrt{\ell_1\ell_2}$ is a norm on $\mathbb{R}^p$.* □

Proof

□

Theorem 1

Because $\hat{\beta}(\lambda) = \arg\min_{\beta}\{L_{PED}(\lambda, \beta)\}$, we have $\frac{\partial L_{PED}(\lambda,\beta)}{\partial \beta_k}\Big|_{\beta=\hat{\beta}(\lambda)} = 0$ for every $k = 1, 2, \ldots, p$ if $\hat{\beta}_k(\lambda) \neq 0$. Thus, we have

$$-\frac{X_k^T[Y - X\hat{\beta}(\lambda)]}{\|Y - X\hat{\beta}(\lambda)\|} + \frac{\lambda}{2}\frac{\frac{\hat{\beta}_k(\lambda)}{\|\hat{\beta}(\lambda)\|}|\hat{\beta}(\lambda)|_1}{\sqrt{\|\hat{\beta}(\lambda)\||\hat{\beta}(\lambda)|_1}} + \frac{\lambda}{2}\frac{sgn\{\hat{\beta}_k(\lambda)\}\|\hat{\beta}(\lambda)\|}{\sqrt{\|\hat{\beta}(\lambda)\||\hat{\beta}(\lambda)|_1}} = 0. \tag{6}$$

If we take $k = i$ and $k = j$, after subtraction, we obtain $\frac{[X_j^T - X_i^T][Y - X\hat{\beta}(\lambda)]}{\|Y - X\hat{\beta}(\lambda)\|} + \frac{\lambda}{2}\frac{[\hat{\beta}_i(\lambda) - \hat{\beta}_j(\lambda)]|\hat{\beta}(\lambda)|_1}{\sqrt{\|\hat{\beta}(\lambda)\|^3|\hat{\beta}(\lambda)|_1}} = 0$ because $\text{sgn}\{\hat{\beta}_i(\lambda)\} = \text{sgn}\{\hat{\beta}_j(\lambda)\}$. *Then, it follows that*

$$\frac{\hat{\beta}_i(\lambda) - \hat{\beta}_j(\lambda)}{\|\hat{\beta}(\lambda)\|} = \frac{2}{\lambda}\frac{\sqrt{\|\hat{\beta}(\lambda)\||\hat{\beta}(\lambda)|_1}}{|\hat{\beta}(\lambda)|_1}[X_j^T - X_i^T]\hat{r}(\lambda), \tag{7}$$

where $\hat{r}(\lambda) = \frac{y - X\hat{\beta}(\lambda)}{\|y - X\hat{\beta}(\lambda)\|}$ and $\|X_j^T - X_i^T\|^2 = 2(1-\rho)$ because X is standardized and $\rho = \cos(\theta_{ij})$. We have $\frac{\sqrt{\|\beta\|\|\beta\|_1}}{\|\beta\|_1} \leq 1$ for any non-zero vector β in $\mathbb{R}^p$ and $|\hat{r}(\lambda)| \leq 1$. Thus, Equation (7) implies that

$$D_\lambda(i, j) \leq \frac{2|\hat{r}(\lambda)|}{\lambda}\|X_i - X_j\| \leq \frac{2}{\lambda}\sqrt{2(1-\rho)} \leq 2\frac{\theta_{ij}}{\lambda}, \tag{8}$$

which proves the grouping effect property for the proposed method. □

Proof

□

Proposition 1

Here, we are going to prove the necessity part of the statement because the sufficiency follows directly from the previous lemma. Let us assume that

$$\hat{\beta}(\lambda) = (\hat{\beta}_1(\lambda), \ldots, \hat{\beta}_{j-1}(\lambda), 0, \hat{\beta}_{j+1}(\lambda) \ldots \hat{\beta}_p(\lambda))^T = \arg\min_{\beta}\{L_{PED}(\lambda, \beta)\}$$

for a given $\lambda > 0$. Here, we can fix λ, and for brevity, we can omit it from notations in the course of this proof. For any $t > 0$, we have

$$\frac{L_{PED}(\hat{\beta}_1, \ldots \hat{\beta}_{j-1}, t, \hat{\beta}_{j+1}, \ldots \hat{\beta}_p) - L_{PED}(\hat{\beta}_1, \ldots \hat{\beta}_{j-1}, 0, \hat{\beta}_{j+1}, \ldots \hat{\beta}_p)}{t} \geq 0.$$

Again, for brevity, we can denote $\hat{\beta}_{t@j} = (\hat{\beta}_1, \ldots \hat{\beta}_{j-1}, t, \hat{\beta}_{j+1}, \ldots \hat{\beta}_p)^T$ and also let $\hat{\theta}_{t@j}$ be the angle between $x_{*,j}$ and $Y - X\hat{\beta}_{t@j}$. With the mean value theorem (Lagrange), there exists $0 < t^* < t$ such that

$$\frac{L_{PED}(\hat{\beta}_{t@j}) - L_{PED}(\hat{\beta}_{0@j})}{t} = -\cos(\hat{\theta}_{t^*@j}) + \lambda \frac{\sqrt{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1} - \sqrt{\|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}}{t}.$$

If we rationalize the numerator of the second fraction in the previous equation, we obtain

$$\frac{L_{PED}(\hat{\beta}_{t@j}) - L_{PED}(\hat{\beta}_{0@j})}{t} = -\cos(\hat{\theta}_{t^*@j}) + \lambda \frac{\frac{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1 - \|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}{t}}{\sqrt{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1} + \sqrt{\|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}},$$

and thus, $\cos(\hat{\theta}_{t^*@j}) \leq \lambda \frac{\frac{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1 - \|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}{t}}{\sqrt{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1} + \sqrt{\|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}}$. Also,

$$\frac{\|\hat{\beta}_{t@j}\||\hat{\beta}_{t@j}|_1 - \|\hat{\beta}_{0@j}\||\hat{\beta}_{0@j}|_1}{t} = |\hat{\beta}_{t@j}|_1 \frac{\|\hat{\beta}_{t@j}\| - \|\hat{\beta}_{0@j}\|}{t} + \|\hat{\beta}_{0@j}\| \frac{|\hat{\beta}_{t@j}|_1 - |\hat{\beta}_{0@j}|_1}{t},$$

and we notice that $\frac{|\hat{\beta}_{t@j}|_1 - |\hat{\beta}_{0@j}|_1}{t} = 1$ for any $t > 0$. Letting $t \to 0$, we obtain $\cos(\hat{\theta}_{0@j}) \leq \frac{\lambda}{2}\sqrt{\frac{\|\hat{\beta}_{0@j}\|}{|\hat{\beta}_{0@j}|_1}} = \frac{\lambda\hat{k}}{2}$.

Analogously, by starting with $t < 0$, we can show that $\cos(\hat{\theta}_{0@j}) \geq -\frac{\lambda}{2}\sqrt{\frac{\|\hat{\beta}_{0@j}\|}{|\hat{\beta}_{0@j}|_1}} = \frac{\lambda\hat{k}}{2}$. □

Proof

□

Proposition 2

By writing the necessary conditions for optimality in the case of problem (2)*, we have that $\text{sgn}(X_j^T(Y - X\hat{\beta})) = \text{sgn}(\frac{\pi}{2} - \hat{\theta}_j)$ and $\frac{\hat{\beta}_j(\lambda)}{\|\hat{\beta}(\lambda)\|} = \hat{k}\left(\frac{2X_j^T(Y - X\hat{\beta})}{\lambda\|Y - X\hat{\beta}\|} - \text{sgn}(\hat{\beta}_j(\lambda))\hat{k}\right)$ if $\hat{\beta}_j(\lambda) \neq 0$. Because $\hat{k} > 0$, we have $\text{sgn}(\hat{\beta}_j) = \text{sgn}(X_j^T(Y - X\hat{\beta})) = \text{sgn}(\cos(\hat{\theta}_j))$.* □

Proof

□

Lemma 3

The proof follows directly from (2) *and* (1). □

We make the observation that if $\hat{\beta}(\lambda)$ is a solution of (2), we have $\cos(\hat{\theta}_j) \leq \frac{\lambda}{2}(\hat{k} + \frac{M}{\hat{k}})$ and therefore $\cos(\hat{\theta}_j) \to 0$ when $\lambda \to 0$ because $M \leq 1$ and $p^{-1/4} \leq \hat{k} \leq 1$.

Theorem 2

The proof follows a similar method to that of Theorem 1 in Belloni et al. (2011). Further details and the proof of the corollary are given in an online supplement of Vasiliu et al. (2017). □

(wileyonlinelibrary.com) https://doi.org/10.1002/sta4.175

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council grant number EP/K022547/1 and Royal Society Wolfson Research Merit Award WM110140.

References

Belloni, A, Chernozhoukov, V & Wang, L (2011), 'Square-root Lasso: pivotal recovery of sparse signals via conic programming', *Biometrika*, **98**(4), 791–806.

Bickel, PJ, Ritov, Y & Tsybakov, AB (2009), 'Simultaneous analysis of Lasso and Dantzig selector', *Annals of Statistics*, **37**(4), 1705–1732.

Bonnans, JF, Gilbert, JC, Lemaréchal, C & Sagastizábal, CA (2006), *Numerical Optimization: Theoretical and Practical Aspects*, *Springer-Verlag*, Berlin.

Borggaard, C & Thodberg, HH (1992), 'Optimal minimal neural interpretation of spectra', *Analytical Chemistry*, **64**, 545–551.

Breiman, L & Friedman, JH (1985), 'Estimating optimal transformations for multiple regression and correlation', *Journal of the American Statistical Association*, **80**, 580.

Browne, WJ, Dryden, IL, Handley, K, Mian, S & Schadendorf, D (2010), 'Mixed effect modelling of proteomic mass spectrometry data using Gaussian mixtures', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 617–633.

Caballero Gaudes, C, Petridou, N, Francis, S, Dryden, IL & Gowland, P (2013), 'Paradigm free mapping with sparse regression automatically detects single-trial fMRI BOLD responses', *Human Brain Mapping*, **34**, 501–518.

Candès, E & Tao, T (2007), 'The Dantzig selector: statistical estimation when p is much larger than n', *Annals of Statistics*, **35**(6), 2313–2351.

Chambers, JM, Cleveland, WS, Kleiner, B & Tukey, PA (1983), *Graphical Methods for Data Analysis*, *Wadsworth*, Belmont, CA.

Draper, NR & Smith, H (1998), *Applied Regression Analysis*, 3rd edn., *Wiley*, New York.

Fan, J & Lv, J (2008), 'Sure independence screening for ultra-high dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 849–911.

Fan, J, Samworth, R & Wu, Y (2009), 'Ultra-high dimensional feature selection: beyond the linear model', *Journal of Machine Learning Research*, **10**, 2013–2038.

Gunst, RF & Mason, RL (1980), *Regression Analysis and Its Application*, *Marcel Dekker*, New York. 370–371.

Harrison, D & Rubinfeld, DL (1978), 'Hedonic housing prices and the demand for clean air', *Journal of Environmental Economics and Management*, **5**, 81–102.

Hastie, TJ, Tibshirani, RJ & Friedman, JH (2008), *The Elements of Statistical Learning (2nd edition)*, *Springer*, New York.

Hoerl, AE & Kennard, RW (1970a), 'Ridge regression: biased estimation for nonorthogonal problems', *Technometrics*, **12**, 55–67.

Hoerl, AE & Kennard, RW (1970b), 'Ridge regression: applications to nonorthogonal problems (Corr: V12 p723)', *Technometrics*, **12**, 69–82.

James, G, Witten, D, Hastie, TJ & Tibshirani, RJ (2013), *An Introduction to Statistical Learning with Applications in R*, *Springer*, New York.

Lewis, AS & Overton, ML (2008), *Nonsmooth optimization via BFGS*, *New York University*, Technical Report.

Lichman, M (2013), *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of, California, School of Information and Computer Science*.

Mian, S, Ugurel, S, Parkinson, E, Schlenzka, I, Dryden, IL, Lancashire, L, Ball, G, Creaser, C, Rees, R & Schadendorf, D (2005), 'Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients', *Journal of Clinical Oncology*, **33**, 5088–5093.

Negahban, SN, Ravikumar, P, Wainwright, MJ & Yu, B (2012), 'A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers', *Statistics Science*, **27**, 538–557.

Nocedal, J (1980), 'Updating quasi-Newton matrices with limited storage', *Mathematics of Computation*, **35**, 773–782.

Quinlan, JR (1993), Combining instance-based and model-based learning, *Proceedings on the Tenth International Conference of Machine Learning in Utgoff, PE (ed.), Morgan Kaufmann*, *San Mateo*, 236–243.

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

Tibshirani, RJ, Saunders, M, Rosset, S, Zhu, J & Knight, K (2005), 'Sparsity and smoothness via the fused Lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 91–108.

Vasiliu, D, Dey T & Dryden IL (2017), *Penalized Euclidean distance regression. arXiv 1405:4578*.

Zou, H & Hastie, TJ (2005), 'Regularization and variable selection via the Elastic Net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.