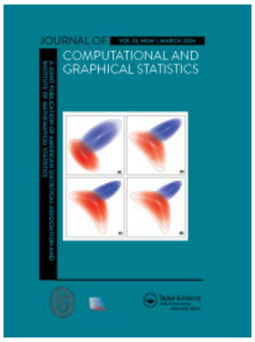

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Improved Pathwise Coordinate Descent for Power Penalties

Maryclare Griffin

To cite this article: Maryclare Griffin (2024) Improved Pathwise Coordinate Descent for Power Penalties, Journal of Computational and Graphical Statistics, 33:1, 310-315, DOI: 10.1080/10618600.2023.2256807

To link to this article: https://doi.org/10.1080/10618600.2023.2256807

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 33, NO. 1, 310–315
https://doi.org/10.1080/10618600.2023.2256807

Taylor & Francis
Taylor & Francis Group

SHORT TECHNICAL NOTE

Improved Pathwise Coordinate Descent for Power Penalties

Maryclare Griffin

Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA

ABSTRACT
Pathwise coordinate descent algorithms have been used to compute entire solution paths for lasso and other penalized regression problems quickly with great success. They improve upon cold start algorithms by solving the problems that make up the solution path sequentially for an ordered set of tuning parameter values, instead of solving each problem separately. However, extending pathwise coordinate descent algorithms to more the general bridge or power family of ℓ_q penalties is challenging. Faster algorithms for computing solution paths for these penalties are needed because ℓ_q penalized regression problems can be nonconvex and especially burdensome to solve. In this article, we show that a reparameterization of ℓ_q penalized regression problems is more amenable to pathwise coordinate descent algorithms. This allows us to improve computation of the mode-thresholding function for ℓ_q penalized regression problems in practice and introduce two separate pathwise algorithms. We show that either pathwise algorithm is faster than the corresponding cold start alternative, and demonstrate that different pathwise algorithms may be more likely to reach better solutions. Supplemental materials for this article are available online.

ARTICLE HISTORY
Received March 2022
Accepted August 2023

KEYWORDS
Coordinate descent; LASSO; Nonconvex optimization; Regularization surface; Sparse regression

1. Introduction

Consider the problem of computing regression coefficients subject to an ℓ_q penalty, sometimes called a bridge or power penalty, which minimizes

$$\frac{1}{2}\left|\left|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2+\lambda\left|\left|\boldsymbol{\beta}\right|\right|_q^q \tag{1}$$

with respect to $\boldsymbol{\beta}$, where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times p$ matrix of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and λ is a tuning parameter (Frank and Friedman 1993). The ℓ_q penalty includes the ℓ_2 ridge, ℓ_0 best subset, and ℓ_1 lasso penalties as special cases. When $q < 1$ the ℓ_q penalty is nonconvex and multiple minimizers of (1) may exist, however ℓ_q penalties are nonetheless valued because they can yield sparser solutions with less bias (Huang, Horowitz, and Ma 2008). It can be difficult to compute a value of $\boldsymbol{\beta}$ that minimizes (1), especially when $q < 1$ (Mazumder, Friedman, and Hastie 2011). This is often magnified by the need to find minimizers of (1) for a collection of values of λ and q, because a single optimal choice of λ and q is rarely known and often data dependent (Griffin and Hoff 2020).

Pathwise coordinate descent algorithms have been one popular approach to overcoming the computational challenges encountered when solving penalized regression problems. Coordinate descent algorithms provide a method for solving (1) for fixed λ and q that compute a solution to (1) by iteratively minimizing with respect to one coordinate of $\boldsymbol{\beta}$ at a time, holding the rest fixed. This corresponds to iteratively computing the mode-thresholding function, $g\left(\lambda, q; b\right)$, which minimizes

$$\frac{1}{2}\left(b-\beta\right)^2+\lambda\left|\beta\right|^q \tag{2}$$

with respect to β, where β corresponds to the coordinate β_j that (1) is being maximized with respect to and the values of b and λ are determined by the data and the remaining coordinates of $\boldsymbol{\beta}$. Pathwise algorithms build on methods that solve (1) for fixed λ and q to solve (1) for a collection of values of λ and q. Although not inextricably linked, pathwise algorithms are often built on a foundation of coordinate descent algorithms.

Coordinate descent algorithms which solve (1) by iteratively solving (2) are provided in in Fu (1998) for $q \geq 1$ and Marjanovic and Solo (2014) for $q < 1$. Marjanovic and Solo (2014) also provide conditions on $\boldsymbol{X}$ that guarantee convergence to a local optimum and verifiable conditions for local optimality of a coordinate descent solution. Fu (1998) provides verifiable optimality conditions for $q \geq 1$. In the case of lasso penalized regression, which corresponds to the special case of (1) when $q = 1$, pathwise coordinate descent methods have been developed (Friedman et al. 2007).

These pathwise coordinate descent algorithms solve (1) for a specific value of the tuning parameter λ^* by finding a value λ_0 that ensures that the minimizing $\boldsymbol{\beta}$ is exactly equal to zero, and then solving (1) along a path of tuning parameter values $\lambda_0, \dots, \lambda^*$ using coordinate descent, using the minimizing $\boldsymbol{\beta}$ for one problem as a starting value for the next. This works well when $q = 1$ because λ_0 is easy to determine from the data, and because mode-thresholding function, $f\left(\lambda; b\right) =$

CONTACT Maryclare Griffin maryclaregri@umass.edu Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

$\text{argmin}_{\beta}\,(b-\beta)^2 + \lambda\,|\beta|$, is nested in λ, that is, if $f\left(\lambda'; b\right) = 0$ and $\epsilon > 0$, then $f\left(\lambda' + \epsilon; b\right) = 0$.

Unfortunately, pathwise coordinate descent algorithms have been more challenging to develop for ℓ_q penalties (Mazumder, Friedman, and Hastie 2011). There are three main challenges that arise in the development of pathwise coordinate descent algorithms for the bridge/power family of penalties, all of which are related to properties of the mode-thresholding function, $g\left(\lambda, q; b\right)$. First, the mode-thresholding function is only available in closed form in special cases. When $q = 1$, $h\left(\lambda, q = 1; b\right) = \text{sign}\,(b)\,(|b| - \lambda)_+$, where $(x)_+ = \max\{x, 0\}$. When $q = 2$, $h\left(\lambda, q = 2; b\right) = (1 + 2\lambda)^{-1}\,b$. Otherwise, a closed-form solution is not available. This means that initializing a pathwise coordinate descent algorithm for fixed values of $q < 1$ by finding a value of the tuning parameter λ_0 that ensures that the solution to (1) is exactly equal to the zero vector can require additional iterative computation, which is possible but can be inconvenient in practice. Second, the mode-thresholding function has two solutions at a single value of b when $q < 1$. Third, the mode-thresholding function is not nested in q for fixed λ, that is, if $g\left(\lambda', q'; b\right) = 0$ and $0 \leq \epsilon < q'$, it may not be true that $g\left(\lambda', q' - \epsilon; b\right) = 0$. This can be observed in the left panel of Figure 1, in which the tuning parameter λ is fixed at 1. For a small range of values of $b \approx 1.5$, the mode-thresholding function returns zero when $q = 0.5$ but not when $q = 0.05$. This is counterintuitive. As q decreases, the corresponding penalty is expected to encourage sparsity more aggressively.

The first two challenges have been addressed by Marjanovic and Solo (2014), who introduced a coordinate descent algorithm for solving (1) for $q < 1$ which includes a method for computing the mode-thesholding function $g\left(\lambda, q, b\right)$, instructions for choosing a single solution to the mode-thresholding function when multiple exist in the context of coordinate descent, and conditions for when the mode-thresholding function $g\left(\lambda, q, b\right) = 0$. In this technical note, we show that a simple reparametrization of (1) eliminates the second challenge for $0 < q < 2$,

$$\frac{1}{2}\left|\left|\boldsymbol{y} - \boldsymbol{X\beta}\right|\right|_2^2 + \left(\frac{\omega^{2-q}}{q}\right)\left|\left|\boldsymbol{\beta}\right|\right|_q^q, \tag{3}$$

in which the tuning parameter λ is replaced by the tuning parameter ω. We emphasize that we use this reparameterization to provide new pathwise coordinate descent algorithms with desirable properties that can make use of existing coordinate descent algorithms for fixed q and ω, as opposed to new coordinate descent algorithms for solving (1) for fixed q and ω.

The right panel of Figure 1 shows the mode-thresholding function corresponding to (3), $h\left(\omega, q; b\right)$, which minimizes

$$\frac{1}{2}\left(b - \beta\right)^2 + \left(\frac{\omega^{2-q}}{q}\right)|\beta|^q. \tag{4}$$

In Figure 1, the mode-thresholding function appears to be nested in q for fixed ω, that is, if $h\left(\omega', q'; b\right) = 0$ and $0 \leq \epsilon < q' \leq 1$, then $h\left(\omega', q' - \epsilon; b\right) = 0$. Furthermore, for fixed ω there appears to be a unique value b^*_ω of b for which all nonzero values of the mode-thresholding function are equal regardless of q and for which nonzero values of the mode-thresholding function are increasing in q for all $b < b^*_\omega$ and decreasing in q for all $b > b^*_\omega$.

In what follows, we derive properties of the mode-thresholding function $h\left(\omega, q; b\right)$. Specifically, we derive a minimum value ω_q which satisfies $h\left(\omega, q; b\right) = 0$ for all $\omega > \omega_q$. We also prove that $h\left(\omega, q; b\right)$ is nested in ω for fixed q and derive b^*_ω, the value of b for which all nonzero values of the mode-thresholding function are equal regardless of q. Last, we prove that sparsity of $h\left(\omega, q; b\right)$ is nested in q for fixed ω that the nonzero values of $h\left(\omega, q; b\right)$ are increasing in q for $b < b^*_\omega$ and decreasing in q for $b > b^*_\omega$. All proofs are provided in an appendix. We then show how this knowledge can be used to improve computation of $h\left(\omega, q; b\right)$ in practice and introduce two different pathwise coordinate descent algorithms for solving (3). One is similar to the pathwise coordinate descent algorithm for solving the lasso penalized regression problem, insofar as it computes a sequence of solutions for fixed q starting from the a value of ω that yields an optimal value of $\boldsymbol{\beta}$ that is exactly equal to the zero vector for $q \leq 1$. The other computes a sequence of solutions for fixed ω starting from $q = 2$. Because (3) is nonconvex when $q < 1$, we compare not only timing of the two algorithms relative to their cold start alternatives and each other, but also the potential for each algorithm to reach a better mode.

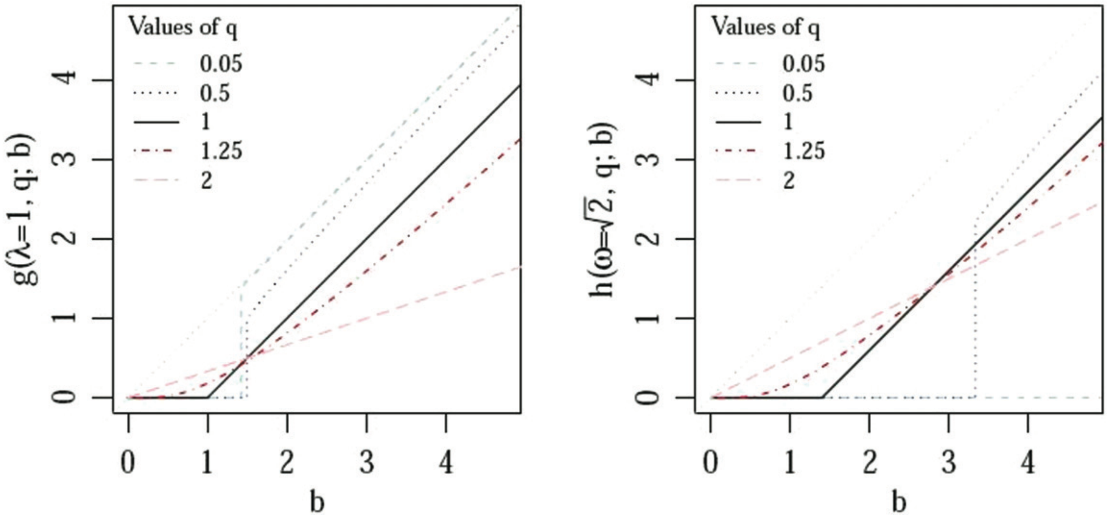

Figure 1. The left panel shows the thresholding function for the parameterization considered in Mazumder, Friedman, and Hastie (2011). The right panel shows the thresholding function for the parameterization given in (4).

2. Properties and New Pathwise Algorithms

Combining the results from Marjanovic and Solo (2014) for $q < 1$ with what is known for $1 \leq q \leq 2$, the mode-thresholding function $h\left(\omega, q; b\right)$ defined in (4) satisfies,

$$h\left(\omega, q; b\right) = \begin{cases} 0 & q \leq 1 \text{ and } |b| < \alpha\left(\omega, q\right) \\ \left\{0, \operatorname{sign}(b)\,\gamma\left(\omega, q\right)\right\} & q \leq 1 \text{ and } |b| = \alpha\left(\omega, q\right) \\ \operatorname{sign}(b)\,\phi\left(\omega, q\right) & q > 1 \text{ or } q \leq 1 \text{ and } |b| > \alpha\left(\omega, q\right), \end{cases} \tag{5}$$

where

$$\gamma\left(\omega, q\right) = \left(\frac{1}{\omega}\right)\left(2\left(\frac{1-q}{q}\right)\right)^{\frac{1}{2-q}} \tag{6}$$

$$\alpha\left(\omega, q\right) = \omega\left(2\left(1-q\right)\right)^{\frac{q-1}{2-q}}\left(2-q\right) q^{\frac{1}{q-2}} \tag{7}$$

and $\phi\left(\omega, q\right) > 0$ is the larger of at most two values that satisfy

$$\phi\left(\omega, q\right) + \omega\left(\frac{\phi\left(\omega, q\right)}{\omega}\right)^{q-1} = |b| \tag{8}$$

when $q > 1$ or $q \leq 1$ and $|b| > \alpha\left(\omega, q\right)$. These properties allow us obtain a condition for ω that ensures $h\left(\omega, q; b\right) = 0$ when $q \leq 1$,

$$\omega > \left(\frac{|b|}{2-q}\right)\left(2\left(1-q\right)\right)^{\frac{1-q}{2-q}} q^{\frac{1}{2-q}}. \tag{9}$$

As suggested by Figure 2, the mode-thresholding function $h\left(\omega, q; b\right)$ is nested in ω and q. Proofs of the following specific claims are provided in an appendix. Focusing on the left panel of Figure 2, we see evidence that sparsity of $h\left(\omega, q; b\right)$ is nested in ω for fixed $q \leq 1$.

Theorem 2.1. For fixed q and $h\left(\omega, q; b\right)$ defined in (5), if $h\left(\omega, q; b\right) = 0$ and $\omega' > \omega$ then $h\left(\omega', q; b\right) = 0$.

The left panel of Figure 2 also suggests that all nonzero mode-thresholding function values for fixed ω are equal when $\omega = |b|/2$, with $\phi\left(\omega = |b|/2, q\right) = |b|/2$.

Theorem 2.2. For $h\left(\omega, q; b\right)$ defined in (5) and q satisfying $q > 1$ or $q \leq 1$ and

$$\left(2/\left(2-q\right)\right)\left(2\left(1-q\right)\right)^{\frac{1-q}{2-q}} q^{\frac{1}{2-q}} > 1, \tag{10}$$

$h\left(\omega = |b|/2, q; b\right) = b/2$.

The left panel of Figure 2 also suggests that nonzero values of $h\left(\omega, q; b\right)$ are nested in ω for fixed q.

Theorem 2.3. For fixed q, if $\left|h\left(\omega, q; b\right)\right| > 0$ and $\omega' < \omega$ then $\left|h\left(\omega', q; b\right)\right| \geq \left|h\left(\omega, q; b\right)\right| > 0$.

Turning to the right panel of Figure 2, we see evidence that sparsity of $h\left(\omega, q; b\right)$ is nested in q for fixed ω.

Theorem 2.4. For fixed ω and $h\left(\omega, q; b\right)$ defined in (5), if $h\left(\omega, q; b\right) = 0$ and $q' < q \leq 1$ then $h\left(\omega, q'; b\right) = 0$.

The right panel of Figure 2 also shows evidence that nonzero values of $h\left(\omega, q; b\right)$ are nested in q for fixed ω.

Theorem 2.5. For fixed $\omega > 0$, define

$$\tilde{q}_{\omega/|b|} = \inf\left\{q : q = 1 \text{ or } q < 1 \text{ and } \left(\frac{1}{2-q}\right) \times \left(2\left(1-q\right)\right)^{\frac{1-q}{2-q}} q^{\frac{1}{2-q}} > \omega/|b|\right\}.$$

If $\omega < |b|/2$ and $q' > q > \tilde{q}_{\omega/|b|}$, then $\left|h\left(\omega, q; b\right)\right| \geq \left|h\left(\omega, q'; b\right)\right| > 0$. If $\omega > |b|/2$ and $q' > q > \tilde{q}_{\omega/|b|}$, then $\left|h\left(\omega, q'; b\right)\right| \geq \left|h\left(\omega, q; b\right)\right| > 0$.

These properties motivate two pathwise coordinate descent algorithms. The first computes a sequence of solutions for fixed

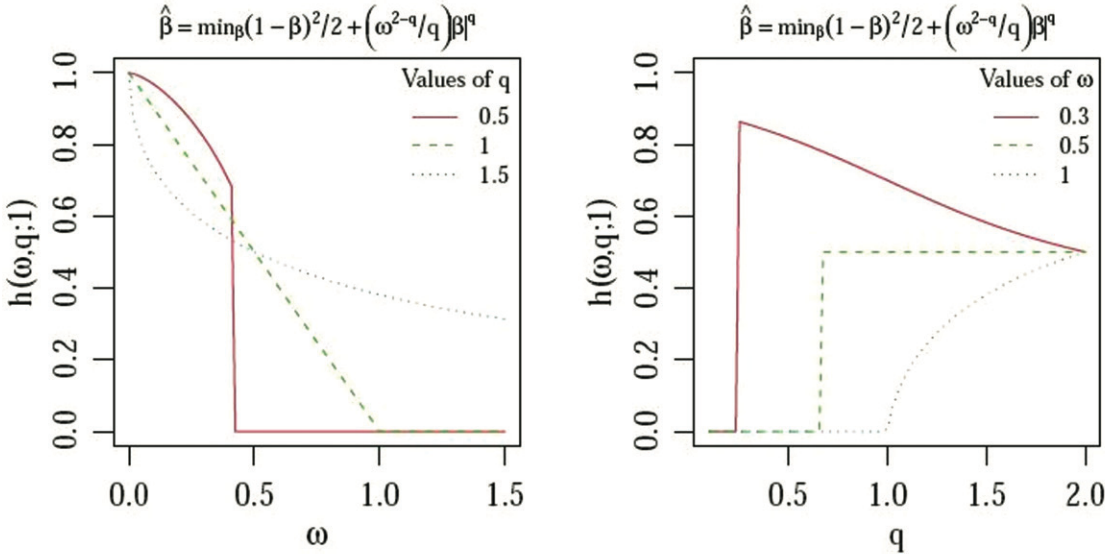

Figure 2. The left panel shows the thresholding function for the parameterization given in (4) as a function of ω for fixed q, the right panel as a function of q for fixed ω.

q and varying values of ω. The second computes a sequence of solutions for fixed ω for varying values of q. We refer to Algorithms 1 and 2 as warm start algorithms, as both repeatedly solve (3) starting from initial values obtained by solving a similar problem previously.

3. Demonstrations

To demonstrate the utility of Algorithms 1 and 2, we consider five simulation settings and applications to five datasets of varying size and structure. For all simulated and real datasets, the response $\boldsymbol{y}$ and covariates $\boldsymbol{X}$ are centered and scaled.

In all simulation settings, we assume $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{z}$, where $\boldsymbol{V}$ and $\boldsymbol{A}$ are $n \times p$ matrix and $p \times p$ matrices, $\boldsymbol{z}$ is a $n \times 1$ noise vector, and nonzero elements of $\boldsymbol{\beta}$, elements of $\boldsymbol{V}$, and elements of $\boldsymbol{z}$ are independent, identically distributed standard normal random variables. Specification of n, p, $\boldsymbol{A}$, and the number of nonzero elements of $\boldsymbol{\beta}$ depends on the simulation setting as described in Table 1. We simulate four datasets per setting.

When implementing Algorithm (1) for fixed q, we consider 20 values of ω that are equally spaced on the log-scale from $\omega_1^{(\min)}$ to 10^{-20}, where $\omega_1^{(\min)}$ is determined from the data. When implementing Algorithm (2) for fixed ω, we consider $k_q = 20$ equally spaced values $\boldsymbol{q} = \{2, \dots, 0.1\}$. For each simulated dataset, we implement Algorithms 1 and 2 and their cold start alternatives for 10 randomly selected unique orderings of the p covariates.

Algorithm 1 Fixed q.

```
1. Based on (9), define
   $$\omega_q^{(min)} = \begin{cases} \max_j \left(\boldsymbol{x}_j'\boldsymbol{x}_j\right)^{\frac{q-1}{2-q}} \left(\frac{\left|\boldsymbol{x}_j'\boldsymbol{y}\right|}{2-q}\right) \left(2\left(1-q\right)\right)^{\frac{1-q}{2-q}} q^{\frac{1}{2-q}} & \text{if } q \leq 1 \\ \max_j \left|\boldsymbol{x}_j'\boldsymbol{y}\right| & \text{otherwise.} \end{cases}$$
   For $q \leq 1$, this is the smallest value of $\omega$ that yields an exactly zero solution for $\boldsymbol{\beta}$.
2. Fix a sequence of $k_\omega$ strictly decreasing $\omega$ values, $\boldsymbol{\omega} = \left\{\omega_1, \dots, \omega_{k_\omega}\right\}$ where $\omega_1 = \omega_q^{(\min)}$. Let $\boldsymbol{B}$ refer to the $p \times k_\omega$ array of solutions.
3. Solve for $\omega = \omega_1$ starting from $\boldsymbol{\beta} = \boldsymbol{0}$, and set the first column of $\boldsymbol{B}$, $\boldsymbol{b}_1$, to the solution.
4. For $l = 2, \dots, k_\omega$, solve for $\omega = \omega_l$ starting from the solution for $\omega_{l-1}$, $\boldsymbol{b}_{l-1}$, and set the $l$th column of $\boldsymbol{B}$, $\boldsymbol{b}_l$, to the solution for $\omega_l$.
5. Return the $p \times k_\omega$ solution array $\boldsymbol{B}$.
```

Algorithm 2 Fixed ω.

```
1. Fix a sequence of $k_q$ decreasing $q$ values $\boldsymbol{q} = \left\{q_1, \dots, q_{k_q}\right\}$, where $q_1 = 2$. Let $\boldsymbol{B}$ refer to the $p \times l$ array of solutions.
2. For $q = q_1$, set the first column of $\boldsymbol{B}$ to the closed-form solution $\boldsymbol{b}_1 = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$.
3. For $l = 2, \dots, k_q$, solve for $q = q_l$ starting from the solution for $q_{l-1}$, $\boldsymbol{b}_{l-1}$, and set the $l$th column of $\boldsymbol{B}$, $\boldsymbol{b}_l$, to the solution for $q_l$.
4. Return the $p \times k_q$ solution array $\boldsymbol{B}$.
```

Figure 3 depicts timing comparisons for Algorithms 1 and 2 relative to cold start alternatives for simulated data. We compare Algorithm 1 to a cold start alternative that minimizes (3) starting from $\boldsymbol{\beta} = \boldsymbol{0}$, and we compare Algorithm 2 to a cold start alternative that minimizes (3) starting from $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}$. In general, both warm start algorithms provide substantial timing gains relative to their cold start alternatives, especially when the number of covariates p is large. The relative gains of Algorithm 1 compared to its cold start alternative are greater than the relative gains of Algorithm 2 compared to its cold start alternative, however Algorithm 2 is often faster than Algorithm 1 when considering time needed to compute solutions for all q and ω, especially when covariates are correlated or the dimension is greater.

Table 2 summarizes the number of observations n, the number of covariates p, and the average absolute correlation between covariates. Citations for where each dataset has previously appeared in the penalized regression literature are also provided.

When implementing Algorithm (1) for fixed q, we consider 20 values of ω that are equally spaced on the log-scale from $\omega_1^{(\min)}$ to 10^{-7}, where $\omega_1^{(\min)}$ is determined from the data. When implementing Algorithm (2) for fixed ω, we consider $k_q = 20$ equally spaced values $\boldsymbol{q} = \{2, \dots, 0.1\}$. For each dataset, we implement Algorithms 1 and 2 and their cold start alternatives for 100 randomly selected unique orderings of the p covariates.

Figure 4 depicts timing comparisons for Algorithms 1 and 2 relative to cold start alternatives. We compare Algorithm 1 to a cold start alternative that minimizes (3) starting from $\boldsymbol{\beta} = \boldsymbol{0}$, and we compare Algorithm 2 to a cold start alternative that minimizes (3) starting from $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}$. Again, both warm start algorithms provide substantial timing gains relative to their cold start alternatives, especially when the number of covariates p is large. Perhaps due to the substantial correlations across covariates in all five of the real datasets, speed advantages of using Algorithm 2 over Algorithm 1 to compute solutions for all q and ω are more pronounced. To assess the extent to which

Table 1. Simulation settings used to demonstrate pathwise coordinate descent Algorithms 1 and 2 for ℓ_q regression. A $p \times p$ identity matrix is denoted by $\boldsymbol{I}_p$.

Simulation	n	p	$\frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{\beta_i=0\}}$	$\boldsymbol{A}$
1	100	1,000	1	$\boldsymbol{I}_p$
2 (Sparse $\boldsymbol{\beta}$)	100	1,000	0.1	$\boldsymbol{I}_p$
3 (Correlated $\boldsymbol{X}$)	100	1,000	1	$0.75\left(\boldsymbol{1}_p\boldsymbol{1}_p'\right)$ $+ (1 - 0.75)\,\boldsymbol{I}_p$
4	500	1,000	1	$\boldsymbol{I}_p$
5	100	2,000	1	$\boldsymbol{I}_p$

Table 2. Features of datasets used to demonstrate pathwise coordinate descent Algorithms 1 and 2 for ℓ_q regression.

Dataset	n	p	$\frac{1}{p(p-1)}\sum_{i \neq j}$ $\left|\text{cor}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right|$	Source
Prostate	97	8	0.295	Tibshirani (1996)
Diabetes	442	64	0.150	Efron et al. (2004)
Glucose	68	72	0.174	Priami and Morine (2015)
Housing	506	104	0.360	Polson, Scott, and Windle (2014)
Motif	287	195	0.641	Bühlmann and van de Geer (2011)

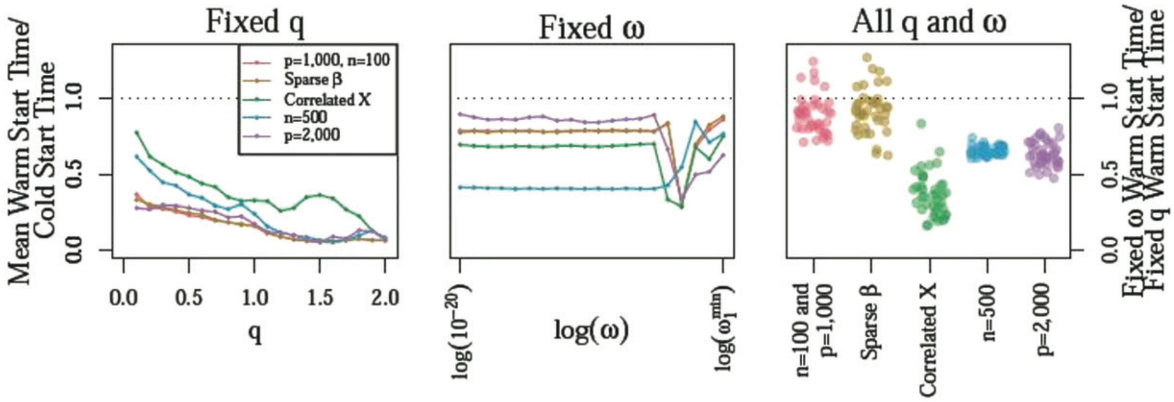

Figure 3. The left and center panels show the mean ratio of time needed to complete Algorithms 1 and 2 compared to their cold start alternatives across 100 randomly selected orderings of the p covariates. The right panel shows the ratio of time needed to complete Algorithm 2 relative to Algorithm 1 for all q and ω for each of the 100 randomly selected orderings of the p covariates.

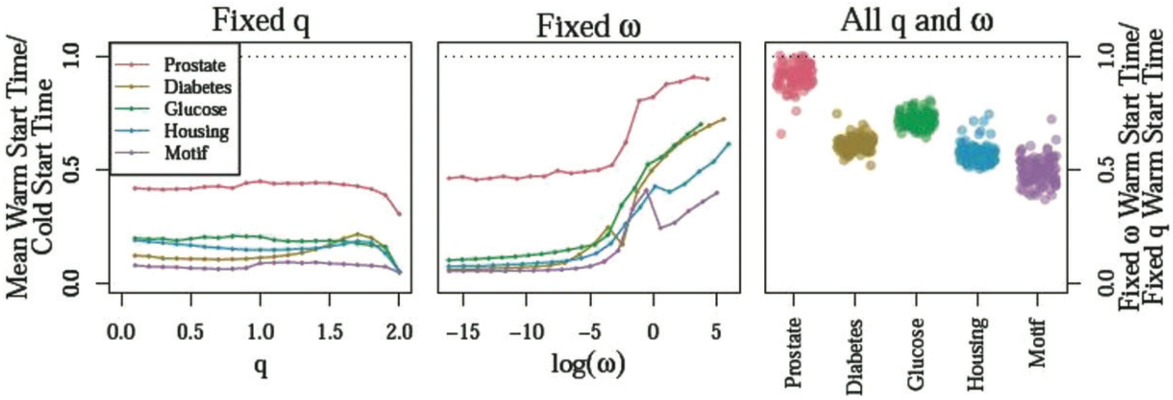

Figure 4. The left and center panels show the mean ratio of time needed to complete Algorithms 1 and 2 compared to their cold start alternatives across 100 randomly selected orderings of the p covariates. The right panel shows the ratio of time needed to complete Algorithm 2 relative to Algorithm 1 for all q and ω for each of the 100 randomly selected orderings of the p covariates.

speed advantages of Algorithm 2 are driven by performance when $q > 1$, which is less often of interest in practice, we have comparisons of a variation of Algorithm 2 that considers only $q \leq 1$ to its cold start alternative and Algorithm 1 in the appendix. We find that speed advantages of Algorithm 2 are only slightly diminished.

A natural question given the favorable timing results for warm start algorithms shown in Figure 4 is whether or not timing gains come at the cost of poorer solutions when $q < 1$ and (1) is nonconvex. Figure 5 compares the rate at which each algorithm reaches an objective value within 10^{-3} of the lowest objective value obtained using the same ordering of covariates. We do not observe that the timing gains associated with Algorithms 1 and 2 relative to their cold start algorithms come at the cost of poorer solutions. Algorithm 1 tends to provide comparable solutions to its cold start alternative for all five datasets. Algorithm 2 tends to provide comparable solutions to its cold start alternative for the prostate and diabetes datasets and better solutions for the glucose, housing, and motif datasets. In general, Algorithm 2 tends to provide the best solutions.

4. Discussion

In this article, we have demonstrated that a new reparameterization of the ℓ_q penalized regression problem is well-suited to pathwise coordinate descent algorithms. There are several potential areas of improvement. One is to explicitly examine which algorithms return solutions that satisfy the conditions provided for local optimality in Marjanovic and Solo (2014) when $q < 1$, instead of considering which algorithms tend to return solutions with the lowest value of the objective function. Another is consideration of the number and spacing of values of ω and q for Algorithms 1 and 2, respectively, which may determine the extent of Algorithms 1 and 2's gains relative to cold start alternatives. A third is consideration of the ordering of covariates in coordinate descent. It is known that flexibility with respect to the ordering of covariates is a specific advantage of coordinate descent methods, and it is possible that gains from Algorithms 1 and 2 relative to their cold start alternatives could be enhanced by certain choices of covariate order such as the ordering used by the active shooting algorithm described in Peng et al. (2009) and the orderings discussed in Chartrand and Yin (2016). A fourth is use of alternative methods for solving (1) for fixed values of q and ω, for example, the local quadratic approximation approach of Fan and Li (2001), the modified Newton-Raphson approach for $1 < q < 2$ described in Fu (1998), or other alternatives reviewed in Chartrand and Yin (2016). Last, further work may consider the choice of a single solution to (1) when multiple exist.

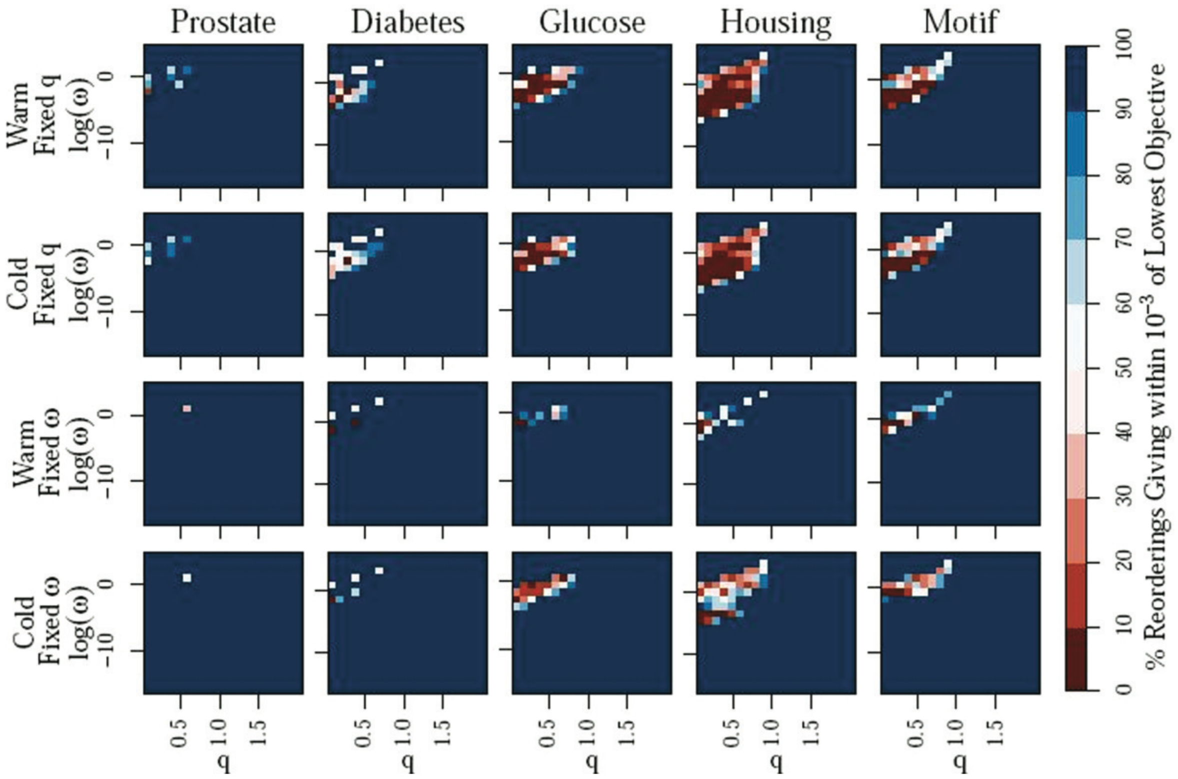

Figure 5. For each dataset, algorithm, and pair of tuning parameter values q and ω, the proportion of random orderings of the covariates where the corresponding algorithm returns an objective function value within 10^{-3} of the lowest objective function value achieved by any of the remaining three algorithms using the same ordering. Warm fixed q corresponds to Algorithm 1, cold fixed q corresponds to Algorithm 1's cold start alternative, warm fixed ω corresponds to Algorithm 2, and cold fixed ω corresponds to Algorithm 2's cold start alternative.

Supplementary Materials

Appendix: Proofs and additional comparisons. (`appx_powcd.pdf`, PDF document)

Code and Data: Code and data for replication. (`JCGS_Submission_Code.zip`, ZIP archive)

Data Availability Statement

Replication code is available at *https://github.com/maryclare/powreg*.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research was supported by NSF grant DMS-2113079.

References

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Berlin: Springer. [313]

Chartrand, R., and Yin, W. (2016), "Nonconvex Sparse Regularization and Splitting Algorithms," in *Splitting Methods in Communication, Imaging, Science, and Engineering*, eds. R. Glowinski, S. J. Osher, and W. Yin, pp. 237–249, Cham: Springer. [314]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [313]

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. [314]

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135. [310]

Friedman, J. H., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302–332. [310]

Fu, W. J. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416. [310,314]

Griffin, M., and Hoff, P. D. (2020), "Testing Sparsity Inducing Penalties," *Journal of Computational and Graphical Statistics*, 29, 1–12. [310]

Huang, J., Horowitz, J. L., and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *The Annals of Statistics*, 36, 587–613. [310]

Marjanovic, G., and Solo, V. (2014), "lq Sparsity Penalized Linear Regression with Cyclic Descent," *IEEE Transactions on Signal Processing*, 62, 1464–1475. [310,311,312,314]

Mazumder, R., Friedman, J. H., and Hastie, T. (2011), "Sparsenet: Coordinate Descent with Nonconvex Penalties," *Journal of the American Statistical Association*, 106, 1125–1138. [310,311]

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009), "Partial Correlation Estimation by Joint Sparse Regression Models," *Journal of the American Statistical Association*, 104, 735–746. [314]

Polson, N. G., Scott, J. G., and Windle, J. (2014), "The Bayesian Bridge," *Journal of the Royal Statistical Society*, Series B, 76, 713–733. [313]

Priami, C., and Morine, M. J. (2015), *Analysis of Biological Systems*, London: Imperial College Press. [313]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [313]