Bicoordinate Descent for the LASSO*

In Song Kim† John Londregan‡ Marc Ratkovic§

Version 0.3 – December 25, 2014

Abstract

We propose an estimator for the LASSO that converges faster than the standard coordinate wise descent algorithm.

Key Words: variable selection, LASSO

Introduction

We offer an improvement to the coordinate wise descent method for estimating the LASSO pioneered by Tibshirani (1996), Fu (1998), and by Friedman, Hastie and Tibshirani (2010a). Our bicoordinate descent method generalizes the one parameter at a time soft thresholding embodied in Fu's "shooting algorithm" by updating the parameter values two at a time. When the regressors are orthogonal our algorithm coincides with Tibshirani's one coordinate at a time soft thresholding, but when the explanators are correlated our algorithm exploits this to more efficiently update the coefficient estimates in pairs. This results in a substantial reduction in the number of passes through the data that the algorithm takes on its path to convergence. The time required by our method for each pass through the data increases relative to unicoordinate descent by much less than the number of passes falls, resulting in an overall improvement in the time to convergence.

Tibshirani (1996) promulgated the LASSO model as a practical sparse estimator. He noted that in the special case of orthogonal regressors a remarkably straightforward solution can be found by

*The proposed methods can be implemented via the open-source statistical software, `bcd: Bicoordinate Descent for the LASSO`, available through the Comprehensive R Archive Network (http://cran.r-project.org/package=bcd).

†Assistant Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, 02139. Email: insong@mit.EDU, URL: http://web.mit.edu/insong/www/

‡Professor of Politics and International Affairs, Woodrow Wilson School, Princeton University, Princeton NJ 08544. Phone: 609-258-4854, Email: jbl@princeton.edu

§Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://www.princeton.edu/~ratkovic

individually soft thresholding each of the estimated coefficients. Fu (1998) developed a "shooting" algorithm that generalizes this approach to any set of regressors–at each pass through the data the algorithm successively updates the parameters one at a time using soft thresholding. Convergence of Fu's algorithm is quick, and Friedman, Hastie and Tibshirani (2010a) make the algorithm even faster by arraying solutions to a sequence of LASSO problems in a trellis, and using each solution as a starting value for the next problem, their glmnet software, which embodies a brace of best programming practices, has defined the computational frontier for the LASSO model.

Our exposition proceeds as follows. The next section introduces our bicoordinate descent algorithm for LASSOed regression, and provides graphical intuition about its workings. Proofs are provided in the appendix. The subsequent section discusses some algorithmic adaptations that accelerate computation. We then turn our attention to the speed with which the respective algorithms converge using various data sets of different sizes. A final section concludes and discusses ongoing directions of research, including the extension of our algorithm to weighted least squares and to generalized linear models such as the logit and the probit. The open-source software, `bcd: Bicoordinate Descent for the LASSO`, for fitting the proposed method is available through the Comprehensive R Archive Network (http://cran.r-project.org/package=bcd).

1 Estimating the LASSO

Starting with data of the form $\left\{Y_i, \{X_{ij}\}_{j=1}^k\right\}_{i=1}^n$ we first center the observations, and normalize the l^2 norm of each of the explanators to equal one, leaving us with: $\left\{y_i, \{x_{ij}\}_{j=1}^k\right\}_{i=1}^n$ satisfying $\sum_{i=1}^n y_i = 0$, and for each $j \in \{1, ..., k\}$ we also have $\sum_{i=1}^n x_{ij} = 0$, and $\sum_{i=1}^n x_{ij}^2 = 1$. If any pairs of explanators are perfectly correlated we arbitrarily remove one element of the perfectly correlated pair, until no perfectly correlated pairs of explanators remain[1].

The LASSO estimator introduced by Tibshirani (1996) is the solution to a problem of the form:

$$P1: \min_{\{\beta_j\}_{j=1}^k} RSS\left(\{\beta_j\}_{j=1}^k | \left\{\{y_i, \{x_{ij}\}_{j=1}^k\right\}_{i=1}^n\right) \text{ subject to } \sum_{j=1}^k |\beta_j| \leq t \tag{1}$$

where:

[1]Of course for each perfectly correlated pair for which at least one element is selected by the LASSO there will in general be a continuum of equivalent solutions to our problem.

$$\mathrm{RSS}\big(\{\beta_j\}_{j=1}^k | \big\{y_i, \{x_{ij}\}_{j=1}^k\big\}_{i=1}^n\big) = \sum_{i=1}^n \Big(y_i - \sum_{j=1}^k \beta_j x_{ij}\Big)^2 \tag{2}$$

Next, recalling that we have culled all of the perfectly correlated observations, let's arrange our data into $C = \lfloor \frac{k}{2} \rfloor$ pairs, with at most one singleton observation, which remains when k is odd.

Now suppose that we take successive passes through the data. At iteration s we turn to each pair of coefficients in turn, taking the others as given at their current values. We seek to minimize the constrained residual sum of squares with respect to $\{\beta_{2c-1}, \beta_{2c}\}$ only, while of course continuing to satisfy the constraint. We can formalize this problem as:

$$\mathrm{P2}_c^s: \quad \min_{\beta_{2c-1}, \beta_{2c}} \mathrm{RSS}_c\big(\beta_{2c-1}, \beta_{2c} | \big\{v_{ic}^s, x_{i,2c-1}, x_{i,2c}\big\}_{i=1}^n\big) \text{ subject to } |\beta_{2c-1}| + |\beta_{2c}| \leq \theta_c^s$$

where:

$$\mathrm{RSS}_c\big(\beta_{2c-1}, \beta_{2c} | \big\{v_{ic}^s, x_{i,2c-1}, x_{i,2c}\big\}_{i=1}^n\big) = \sum_{i=1}^n (v_{ic}^s - \beta_{2c-1} x_{i,2c-1} - \beta_{2c} x_{i,2c})^2$$

and:

$$v_{ic}^s = \Big(y_i - \sum_{j<2c-1} \beta_j^{s,\mathrm{lasso}} x_{ij} - \sum_{2c<j} \beta_j^{s-1,\mathrm{lasso}} x_{ij}\Big) \quad \text{while} \quad \theta_c^s = t - \sum_{j<2c-1} |\beta_j^{s,\mathrm{lasso}}| - \sum_{2c<j} |\beta_j^{s-1,\mathrm{lasso}}|$$

We denote the solutions to $\mathrm{P2}_c^s$ by $(\beta_{2c-1}^{s,\mathrm{lasso}}, \beta_{2c}^{s,\mathrm{lasso}})$.

Finally, if k is odd, there remains a singleton observation that is not encompassed by any of the pairs. Define:

$$\mathrm{P3}^s: \quad \min_{\beta_k} \sum_{i=1}^n \Big(v_{ik}^s - \beta_k x_{i,k}\Big)^2 \text{ subject to } |\beta_k| \leq \theta_p^s$$

where:

$$v_{ik}^s = \Big(y_i - \sum_{j<k} \beta_j^{s,\mathrm{lasso}} x_{ij}\Big) \text{ and } \theta_k^s = t - \sum_{j<k} |\beta_j^{s,\mathrm{lasso}}|$$

and we denote the solutions to $\mathrm{P3}^s$ by $\beta_k^{s,\mathrm{lasso}}$.

The Bicoordinate Descent Algorithm

Our algorithm for calculating $(\beta_{2c-1}^{s,\mathrm{lasso}}, \beta_{2c}^{s,\mathrm{lasso}})$ proceeds as follows. Let $(\beta_{2c-1}^{s,\mathrm{ols}}, \beta_{2c}^{s,\mathrm{ols}})$ solve:

$$\mathsf{POLS}_c^s: \quad \min_{\beta^*_{2c-1},\beta^*_{2c}} \mathsf{RSS}_c\left(\beta^*_{2c-1},\beta^*_{2c}|\left\{v^s_{ic},x_{i,2c-1},x_{i,2c}\right\}_{i=1}^n\right)$$

so:

$$\begin{pmatrix}\beta^{s,ols}_{2c-1}\\ \beta^{s,ols}_{2c}\end{pmatrix} = \frac{1}{1-R_c^2}\begin{pmatrix}\sum_{i=1}^n v^s_{ic}(x_{i,2c-1}-R_c x_{i,2c})\\ \sum_{i=1}^n v^s_{ic}(x_{i,2c}-R_c x_{i,2c-1})\end{pmatrix}$$

where:

$$R_c = \sum_{i=1}^n x_{i,2c-1}x_{i,2c} \tag{3}$$

Next define:

$$R_c^{s*} = sign(\beta^{s,ols}_{2c-1}) \times sign(\beta^{s,ols}_{2c}) \times R_c \tag{4}$$

We let λ denote the Lagrange multiplier associated with the constraint in $\mathsf{P1}$. We will treat this as a "tuning parameter" shared by all of the $\mathsf{P2}_c^s$.

Couched in terms of λ, when:

$$\frac{\lambda}{2(1+R_c^{s*})} < min\{|\beta^{s,ols}_{2c-1}|,|\beta^{s,ols}_{2c}|\} \tag{5}$$

our estimates are calculated as:

$$\begin{pmatrix}\beta^{s,lasso}_{2c-1}\\ \beta^{s,lasso}_{2c}\end{pmatrix} = \begin{pmatrix}sign(\beta^{s,ols}_{2c-1})\left(|\beta^{s,ols}_{2c-1}|-\frac{\lambda}{2(1+R_c^{s*})}\right)\\ sign(\beta^{s,ols}_{2c})\left(|\beta^{s,ols}_{2c}|-\frac{\lambda}{2(1+R_c^{s*})}\right)\end{pmatrix} \tag{6}$$

When condition (5) fails, but

$$|\beta^{s,ols}_{2c-1}| > |\beta^{s,ols}_{2c}| \tag{7}$$

then the update step for $(\beta^{s,lasso}_{2c-1},\beta^{s,lasso}_{2c})$ is:

$$\left(\beta^{s,lasso}_{2c-1},\beta^{s,lasso}_{2c}\right) = sign(\beta^{s,ols}_{2c-1})max\left\{|\beta^{s,ols}_{2c-1}|+R^{*s}_{2c}|\beta^{s,ols}_{2c}|-\frac{\lambda}{2},0\right\} \tag{8}$$

whereas if (5) fails but the inequality in condition (7) is reversed, then:

$$\left(\beta^{s,lasso}_{2c-1},\beta^{s,lasso}_{2c}\right) = sign(\beta^{s,ols}_{2c})max\left\{0,|\beta^{s,ols}_{2c}|+R^{*s}_{2c}|\beta^{s,ols}_{2c-1}|-\frac{\lambda}{2}\right\} \tag{9}$$

Notice that when $R_c = 0$ the bicoordinate descent algorithm coincides with the soft thresholding embodied in the "shooting" algorithm of Fu (1998).

Of course, the solution to $P3^s$ is simply given by the soft thresholding result returned by Fu's algorithm:

$$\beta_p^{s,lasso} = sign(\beta_p^{s,ols}) max \left\{ |\beta_p^{s,ols}| - \frac{\lambda}{2}, 0 \right\} \tag{10}$$

where:

$$\beta_p^{s,ols} = \sum_{i=1}^{n} v_{ic}^s x_{ik}$$

Why it Works

Let's take a closer look at the objective function for $P2_c^s$. First it's useful to define a few terms. For comparison let's start with the sum of squared errors corresponding to the solution to $POLS_c^s$:

$$sse_0^{c,s} = \sum_{i=1}^{n} \left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c} \right)^2 \tag{11}$$

Next consider the following quadratic function $Q(\beta_{2c-1} - \beta_{2c-1}^{s,ols}, \beta_{2c} - \beta_{2c}^{s,ols}, R_c)$ of the correlation R_c between x_{2C-1} and x_{2C} that was defined in (3) and of the distance between the coefficients $(\beta_{2c-1}, \beta_{2c})$ and the OLS coefficients $(\beta_{2c-1}^{OLS}, \beta_{2c}^{OLS})$ with:

$$Q(\beta_{2c-1} - \beta_{2c-1}^{s,ols}, \beta_{2c} - \beta_{2c}^{s,ols}, R_c) = (\beta_{2c-1} - \beta_{2c-1}^{s,ols}, \beta_{2c} - \beta_{2c}^{s,ols}) \begin{pmatrix} 1 & R_c \\ R_c & 1 \end{pmatrix} \begin{pmatrix} \beta_{2c-1} - \beta_{2c-1}^{s,ols} \\ \beta_{2c} - \beta_{2c}^{s,ols} \end{pmatrix} \tag{12}$$

It turns out that we can reconceive the objective function for $P2_c^s$ in terms of Q. We state this formally as:

Lemma 1: $RSS_c\left(\beta_{2c-1}, \beta_{2c} | \left\{ v_{ic}^s, x_{i,2c-1}, x_{i,2c} \right\}_{i=1}^n\right) = sse_0^{c,s} + Q(\beta_{2c-1} - \beta_{2c-1}^{s,ols}, \beta_{2c} - \beta_{2c}^{s,ols}, R_c)$

Proof of Lemma 1: See the appendix.

Notice that $sse_0^{c,s}$ is constant with respect to β_{2c-1} and β_{2c}, so Lemma 1 allows us to reformulate $P2_c^s$ as a quadratic programming problem:

$$P2_c^{s\prime}: \min_{\beta_{2c-1}, \beta_{2c}} Q(\beta_{2c-1} - \beta_{2c-1}^{s,ols}, \beta_{2c} - \beta_{2c}^{s,ols}, R_c) \quad \text{subject to} \quad |\beta_{2c-1}| + |\beta_{2c}| \leq \theta_c^s$$

The constraint is in the form of a diamond, while the level sets of the objective function for $P2_c^{s\prime}$ are ellipses. This is illustrated in the lefthand panel of figure 1, where the hollow dot corresponds to $(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols})$ while $(\beta_{2c-1}^{s,lasso}, \beta_{2c}^{s,lasso})$ is represented by the solid dot.

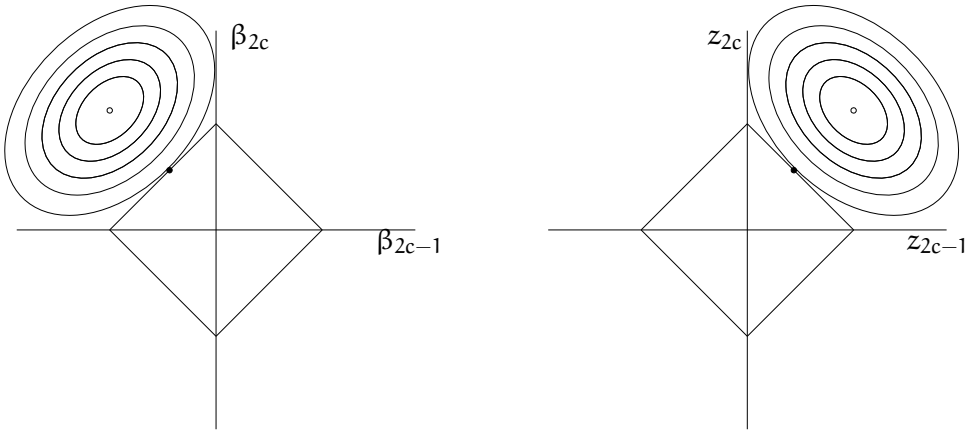

Figure 1: Optimization $P2_c^{s\prime}$ *vs* PZ

There is an isomorphic relationship amongst solutions in distinct quadrants. To see this, define $\delta_j^s \equiv \mathtt{sign}(\beta_j^{s,ols})$ and let $z_j = \delta_j \beta_j$, and then rewrite problem $P2_c^{s\prime}$ as:

$$\min_{z_{2c-1}, z_{2c}} Q\big(\delta_{2c-1}^s z_{2c-1} - \delta_{2c-1}^s |\beta_{2c-1}^{s,ols}|, \delta_{2c}^s z_{2c} - \delta_{2c}^s |\beta_{2c}^{s,ols}|, R_c\big) \quad \text{subject to} \quad |z_{2c-1}| + |z_{2c}| \leq \theta_c^s$$

recalling our definition of R_c^{s*} from expression (4) this can be reexpressed as:

$$PZ: \min_{z_{2c-1}, z_{2c}} Q\big(z_{2c-1} - |\beta_{2c-1}^{s,ols}|, z_{2c} - |\beta_{2c}^{s,ols}|, R_c^{s*}\big) \text{ subject to } |z_{2c-1}| + |z_{2c}| \leq \theta_c^s$$

If $(\hat{z}_{2c-1}^s, \hat{z}_{2c}^s)$ solve PZ then

$$(\beta_{2c-1}^{s,lasso}, \beta_{2c}^{s,lasso}) = (\delta_{2c-1}^s \hat{z}_{2c-1}^s, \delta_{2c}^s \hat{z}_{2c}^s) \tag{13}$$

is a solution to $P2_c^{s\prime}$. The righthand panel of figure 1 depicts the reformulation of $P2_c^{s\prime}$ as PZ. The orientation of the ellipse shifts with the translation to the first quadrant, this corresponds to the change from R_c to R_c^{s*}. The hollow dot in the right panel corresponds to $(|\beta_{2c-1}^{s,ols}|, |\beta_{2c}^{s,ols}|)$ whereas the solid dot indicates the solution values (z_{2c-1}^s, z_{2c}^s) for PZ.

It turns out that the solution to PZ are non-negative. In fact, if the constraint binds the solution is to be found along the first quadrant simplex S:

$$S = \left\{(z_{2c-1}, z_{2c}) | z_{2c-1} + z_{2c} = \theta_c^s \text{ and } z_{2c-1} \geq 0 \text{ and } z_{2c} \geq 0\right\} \tag{14}$$

We state this important result as:

Lemma 2: The solutions to PZ satisfy $\hat{z}_{2c-1} \geq 0$ and $\hat{z}_{2c} \geq 0$, while if $\theta_c^s \leq |\beta_{2c-1}^{OLS}| + |\beta_{2c}^{OLS}|$, then $(\hat{z}_{2c-1}, \hat{z}_{2c}) \in S$.

Proof: See the appendix.

Now let's take a graphical approach to the solution. Consider the objective function:

$$Q\left(z_{2c-1} - |\beta_{2c-1}^{s,ols}|, z_{2c} - |\beta_{2c}^{s,ols}|, R_c^{s*}\right) \tag{15}$$

for PZ. The lefthand panel of figure 2 shows the level curves for Q. At an interior solution for (z_{2c-1}, z_{2c}) the highest level curve that makes contact with the constraint will be tangent to S, and so it will have the same slope, -1, as the simplex, several points at which the slope of a level curve matches -1 are depicted in the lefthand panel of figure 2. The level curve slopes are given by:

$$\frac{dz_{2c}}{dz_{2c-1}} = -\frac{\frac{\partial Q}{\partial z_{2c-1}}}{\frac{\partial Q}{\partial z_{2c}}} \tag{16}$$

Setting this slope to -1 and solving we recover the locus of points at which the level curves of Q share the same slope as S, see the central panel of figure 2:

$$z_{2c} = |\beta_{2c}^{s,ols}| - |\beta_{2c-1}^{s,ols}| + z_{2c-1} \tag{17}$$

Putting this formally, we have:

Lemma 3: The locus of points at which the level curves of Q share the same slope as S is given by (17).

Proof: See the appendix.

If the line (17) intersects S we have a tangency solution for PZ, such a solution is depicted in the righthand panel of figure 2, where it corresponds to the solid dot whose coordinates are given by:

$$(z_{2c-1}, z_{2c}) = \left(\frac{\theta_c^s + |\beta_{2c-1}^{s,ols}| - |\beta_{2c}^{s,ols}|}{2}, \frac{\theta_c^s + |\beta_{2c}^{s,ols}| - |\beta_{2c-1}^{s,ols}|}{2}\right) \tag{18}$$

A tangency solution will only exist if the locus of tangencies intersects S, and expression (17) tells us that this set of tangencies always corresponds to a line with slope 1 passing through $\left(|\beta_{2c-1}|, |\beta_{2c}|\right)$.

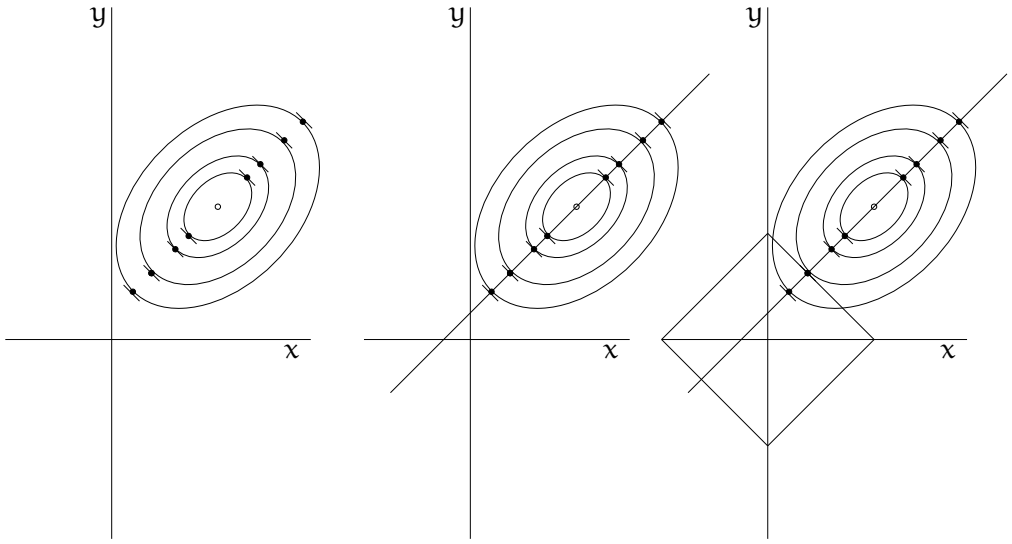

Figure 2: Right: Tangencies Center: Locus of Tangencies Left: Interior Solution

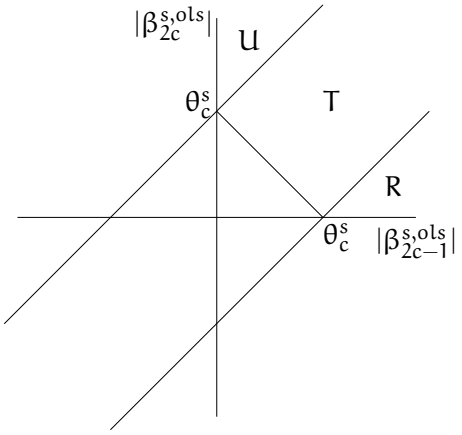

Figure 3: Solutions relative to θ_c^s

This tells us that that whenever $(\beta_{2c-1}, \beta_{2c})$ lie in the region of figure 3 that is marked T in figure 3 southeast of the line through $(|\beta_{2c-1}|, |\beta_{2c}|) = (0, \theta_c^s)$ with slope equal to one:

$$z_{2c} = \theta_c^s + z_{2c-1} \tag{19}$$

and northwest of the line with unit slope that passes through $(|\beta_{2c-1}^{s,ols}|, |\beta_{2c}^{s,ols}|) = (\theta_c^s, 0)$:

$$z_{2c} = -\theta_c^s + z_{2c-1} \tag{20}$$

we will have a tangency solution. Combining expressions (19) and (20) we have a tangency solution given by (18) whenever:

$$\left| |\beta_{2c}^{s,ols}| - |\beta_{2c-1}^{s,ols}| \right| \leq \theta_c^s \tag{21}$$

In contrast, if $|\beta_{2c-1}^{s,ols}|$, and $|\beta_{2c}^{s,ols}|$ lie outside region T in figure 3, and so fail to satisfy condition (21), then Lemma 3 implies we cannot have a tangency solution.

If we have a $(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols})$ pair above the line (19), in the region marked U in figure 3, so that:

$$|\beta_{2c}^{s,ols}| - |\beta_{2c-1}^{s,ols}| \geq \theta_c^s \tag{22}$$

then we cannot have a first quadrant tangency with S. However, the only remaining alternatives are a solution at the upper corner of the constraint set $(0, \theta_c^s)$ and a solution at the righthand corner, $(\theta_c^s, 0)$. It is straightforward to show that in such a case the upper corner of the constraint set:

$$(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols}) = (0, \theta_c^s) \tag{23}$$

provides a better solution. Likewise, if $(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols})$ lie below (20), in the region of figure 3 marked R, so that:

$$|\beta_{2c}^{s,ols}| - |\beta_{2c-1}^{s,ols}| \geq \theta_c^s \tag{24}$$

the solution comes at the right corner:

$$(z_{2c-1}, z_{2c}) = (\theta_c^s, 0) \tag{25}$$

These results on corner solutions are an immediate consequence of the following lemma:

Lemma 4: whenever $\theta_c^s > 0$

$$sign\left(Q(\theta_c^s - |\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|, R_c^*) - Q(-|\beta_{2c-1}^{s,ols}|, \theta_c^s - |\beta_{2c}^{s,ols}|, R_c^*)\right) = sign\left(|\beta_{2c-1}^{s,ols}| - |\beta_{2c}^{s,ols}|\right)$$

Proof: See the appendix[2].

It follows that:

[2]Notice that the case in which $\theta_c^s = 0$ is trivial, as the only possible solution is $(\hat{z}_{2c-1}, \hat{z}_{2c}) = (0, 0)$ in which case distinctions among tangencies and various corner solutions are vacuous.

Corollary A: $(|\beta^{s,ols}_{2c-1}|, |\beta^{s,ols}_{2c}|) \in U$

implies

$$Q(-|\beta^{s,ols}_{2c-1}|, \theta^s_c - |\beta^{s,ols}_{2c}|, R^*_c) > Q(\theta^s_c - |\beta^{s,ols}_{2c-1}|, -|\beta^{s,ols}_{2c}|, R^*_c)$$

Proof: By the definition of U, $(|\beta^{s,ols}_{2c-1}|, |\beta^{s,ols}_{2c}|) \in U$ implies $|\beta^{s,ols}_{2c}| > |\beta^{s,ols}_{2c-1}| + \theta > |\beta^{s,ols}_{2c-1}|$, and so by Lemma 4 we have $Q(-|\beta^{s,ols}_{2c-1}|, \theta^s_c - |\beta^{s,ols}_{2c}|, R^*_c) > Q(\theta^s_c - |\beta^{s,ols}_{2c-1}|, -|\beta^{s,ols}_{2c}|, R^*_c)$. □

Corollary B: $(|\beta^{s,ols}_{2c-1}|, |\beta^{s,ols}_{2c}|) \in R$

implies

$$Q(\theta^s_c - |\beta^{s,ols}_{2c-1}|, -|\beta^{s,ols}_{2c}|, R^*_c) > Q(-|\beta^{s,ols}_{2c-1}|, \theta^s_c - |\beta^{s,ols}_{2c}|, R^*_c)$$

The proof of Corollary B is completely analogous.

It remains for us to solve for θ^s_c.

Given that the constraint is binding, which it will be when $\lambda > 0$, we will have $(z_{2c-1}, z_{2c}) \in S$, and we can reposit PZ as:

$$PZ': \quad \min_{z_{2c-1}, z_{2c}} \quad Q\left(z_{2c-1} - |\beta^{s,ols}_{2c-1}|, z_{2c} - |\beta^{s,ols}_{2c}|, R^{s*}_c\right)$$

$$\text{subject to} \quad z_{2c-1} + z_{2c} \leq \theta^s_c$$
$$z_{2c-1} \geq 0$$
$$z_{2c} \geq 0$$

Formulating the Lagrangian we have:

$$\min_{z_{2c-1}, z_{2c}} L = Q\big(z_{2c-1} - |\beta^{s,ols}_{2c-1}|, z_{2c} - |\beta^{s,ols}_{2c}|, R^{s*}_c\big) + \lambda\left(z_{2c-1} + z_{2c} - \theta^s_c\right) - \mu_{2c-1} z_{2c-1} - \mu_{2c} z_{2c} \quad (26)$$

Now let's consider the possible solutions.

Lemma 5: At an interior solution to (26) with with both $z_{2c-1} > 0$ and $z_{2c} > 0$ we have:

$$\theta^s_c = |\beta^{s,ols}_{2c-1}| + |\beta^{s,ols}_{2c}| - \frac{\lambda}{1 + R_c} \quad (27)$$

Proof: See the Appendix.

Substituting from (27) into (21) and rearranging terms we have our conditions for a tangency solution in terms of $|\beta_{2c-1}^{s,ols}|$, $|\beta_{2c}^{s,ols}|$, and λ:

$$\frac{\lambda}{2(1+R_c)} \leq \min\left\{|\beta_{2c-1}^{s,ols}|, |\beta_{2c}^{s,ols}|\right\} \tag{28}$$

when (28) is satisfied we can substitute from (27) into (18) to obtain our tangency solution:

$$(z_{2c-1}, z_{2c}) = \left(\beta_{2c-1}^{s,ols} - \frac{\lambda}{2(1+R_c^{s*})}, \beta_{2c}^{s,ols} - \frac{\lambda}{2(1+R_c^{s*})}\right) \tag{29}$$

The lefthand panel of figure (4) depicts the solutions when $R_c^{s*} > 0$, while the right hand panel shows the case of $R_c^{s*} < 0$. In each figure, the region marked T, for "tangency", corresponds to condition (28). Notice that for a given value of λ this area is more extensive when $R_c^{s*} > 0$, as shown in the left panel, than it is for negatively correlated pairs of regressors, as depicted in the righthand panel.

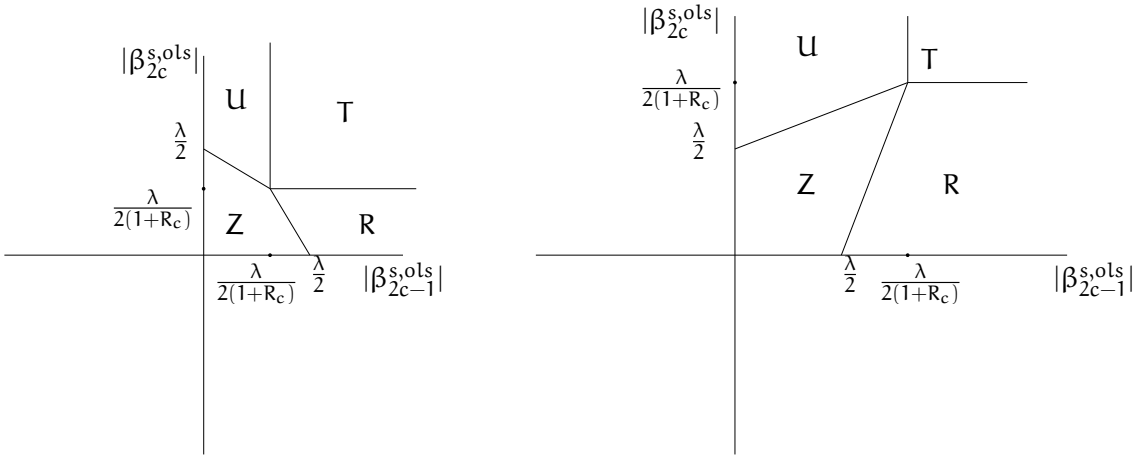

Figure 4: Left: Solutions with $R_c^{s*} > 0$, Right: Solutions with $R_c^{s*} < 0$

Now suppose we have a corner solution with $z_{2c-1} > 0$ but $z_{2c} = 0$.

Lemma 6: At a corner solution to (26) with with $z_{2c-1} > 0$ but $z_{2c} = 0$ we have:

$$\theta_c^s = |\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}| - \frac{\lambda}{2} \tag{30}$$

Proof: See the Appendix.

Of course, this only works provided $\theta_c^s \geq 0$, that is, if:

$$\frac{\lambda}{2} \leq |\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}| \tag{31}$$

Substituting θ_c^s from (30) into $\frac{\partial L}{\partial z_{2c}} \geq 0$ we have:

$$\begin{aligned}
\frac{\partial L}{\partial z_{2c}} &= 2R_c^{s*}(z_{2c} - |\beta_{2c-1}^{s,ols}|) + 2(0 - |\beta_{2c}^{s,ols}|) + \lambda \\
&= 2R_c^{s*}(|\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}| - \frac{\lambda}{2} - |\beta_{2c-1}^{s,ols}|) + 2(0 - |\beta_{2c}^{s,ols}|) + \lambda \\
&= -2|\beta_{2c}^{s,ols}| + 2R_c^{s*2}|\beta_{2c}^{s,ols}| + \lambda(1 - R_c^{s*}) \geq 0
\end{aligned}$$

that is, we need:

$$\frac{\lambda}{2} \geq (1 + R_c^*)|\beta_{2c}^{s,ols}| \tag{32}$$

Combining conditions (31) and (32), we have:

$$(1 + R_c^*)|\beta_{2c}^{s,ols}| \leq \frac{\lambda}{2} \leq |\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}|$$

The set of $(|\beta_{2c-1}^{s,ols}|, |\beta_{2c}^{s,ols}|)$ pairs satisfying this condition corresponds to the region labeled R in figure 4. This region is larger when $R_c^{s*} < 0$, as shown in the right hand panel, than it is when the regressors are positively correlated–the bicoordinate descent LASSO update is more likely to eliminate one of the coefficients at the update step when the correlation between the regressors is negative.

Substituting from (30) into (25) we have:

$$(z_{2c-1}, z_{2c}) = \left(|\beta_{2c}^{s,ols}| + R_c^{s*}|\beta_{2c-1}^{s,ols}| - \frac{\lambda}{2}, 0\right) \tag{33}$$

Likewise, we have a solution at the top corner, with:

$$(z_{2c-1}, z_{2c}) = \left(0, |\beta_{2c}^{s,ols}| + R_c^{s*}|\beta_{2c-1}^{s,ols}| - \frac{\lambda}{2}\right) \tag{34}$$

provided:

$$(1 + R_c^*)|\beta_{2c-1}^{s,ols}| \leq \frac{\lambda}{2} \leq |\beta_{2c}^{s,ols}| + R_c^{s*}|\beta_{2c-1}^{s,ols}|$$

Notice that when $R_c^{s*} < 0$ a wider range of parameter estimates results in one parameter, as in regions R and U, or both coefficients, corresponding to region Z, being updated to zero, see the

lefthand panel of figure 4, than in the case shown in the left panel of $R_c^{s*} > 0$. In either case, with $R_c^{s*} \neq 0$ at each pass through the data the bicoordinate descent algorithm allocates slack across the variables more efficiently than does unicoordinate descent, while in the "knife's edge" case of $R_c^{s*} = 0$ unicoordinate and bicoordinate descent update identically conditional on the remaining parameter estimates.

2 Computational Mechanics

The payoff to our algorithm is the speed with which it computes the LASSO estimates. While bicoordinate descent provides savings in the number of passes to be taken through the data, we need also to be abstemious in the computations required at each pass through the data set. We highlight three areas in which we have attempted to apply best practice programming.

Warm Starts

Firstly, glmnet Friedman, Hastie and Tibshirani (2010b) takes advantage of "warm starts." FHT first find the smallest value for λ that will still set all of the coefficients equal to zero. Their algorithm descends from this value of λ in a sequence of steps, each of which takes its predecessor as a source of a starting value.

We emulate this approach. Let $r_{y,j}$ be given by:

$$r_{y,j} = \sum_{i=1}^{n} y_i x_{j,i}$$

Now define:

$$\lambda^{max} \equiv 2max\{r_{y,j}\}_{j=1}^{p}$$

Next we choose a multiple ϵ of λ_{max} to define the smallest λ value we will consider, $\lambda_{min} = \epsilon\lambda_{max}$. Next we choose a number of "cross pieces", M, for the trellis. Finally, we construct a "shrinkage factor" $\sigma = \epsilon^{-\frac{1}{M}}$ such that $\lambda_{min} = \sigma^M\lambda_{max}$. At each iteration we shrink λ from it's previous value: $\lambda_m = \sigma\lambda_{m-1}$. We then start our calculations with lagged values for $\vec{\beta}$ of $\vec{\beta}_1 = \vec{\beta}_2 = \vec{0}$. Our starting value for round $m \in \{1, ..., M\}$ of our descent to the next cross piece of the trellis is:

$$\vec{\beta}_m^{start} = (1 + \sigma)\vec{\beta}_{m-1} - \sigma\vec{\beta}_{m-2}$$

At each iteration we then update the first and second lags of $\vec{\beta}$. We find that these interpolated "warm starts" provide more advantageous starting values than do the unalloyed elements of $\vec{\beta}_1$.

Sufficient Statistics

Our algorithm calls for us to calculate $(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols})$ at each iteration step. While these calculations depend on the *status quo* values for the coefficients, they also rely on various cross products from the data. We eschew recalculation of the latter.

Let $r_{j,j'}$ be defined analogously with $r_{y,j}$:

$$r_{j,j'} = \sum_{i=1}^{n} x_{j,i} x_{j',i}$$

Notice that in this notation $R_c \equiv r_{2c-1,2c}$.

To be comprehensive, let's suppose there are $k = 2k^* + 1$ explanators. The case of an even number is yet easier. Now formulate the $k \times (k+1)$ matrix S. For $c \leq k^*$, we'll denote row $2c-1$ of S, as $\vec{s}'_{2c-1}$. It's elements are:

$$S_{2c-1,j} = \begin{cases} \frac{-r_{j,2c-1} + r_{2c,2c-1} r_{j,2c}}{1 - r_c^{2c,2c-1}} & j \notin \{2c-1, 2c, k+1\} \\ 0 & j \in \{2c-1, 2c\} \\ \frac{r_{y,2c-1} - r_{2c,2c-1} r_{y,2c}}{1 - r_{2c,2c-1}^2} & j = k+1 \end{cases} \tag{35}$$

Likewise, the elements of row $2c$ of S, $\vec{s}'_{2c}$, are:

$$S_{2c,j} = \begin{cases} \frac{-r_{j,2c} + r_{2c,2c-1} r_{j,2c-1}}{1 - r_c^{2c,2c-1}} & j \notin \{2c-1, 2c, k+1\} \\ 0 & j \in \{2c-1, 2c\} \\ \frac{r_{y,2c} - r_{2c,2c-1} r_{y,2c-1}}{1 - r_{2c,2c-1}^2} & j = k+1 \end{cases} \tag{36}$$

the k^{th} and final row of S, $\vec{s}'_k$, is:

$$S_{k,j} = \begin{cases} -r_{j,k} & 1 \leq j < k \\ 0 & j = k \\ r_{y,k} & j = k+1 \end{cases} \tag{37}$$

Starting from the initial $(k+1) \times 1$ vector $\vec{\alpha}^{s,c,ols}$, where:

$$\alpha_j^{s,c,ols} = \begin{cases} \beta_j^{s,ols} & j \leq 2c-2 \\ \beta_j^{s-1,ols} & 2c-1 < j \leq k \\ 1 & j = k+1 \end{cases} \tag{38}$$

we update $(\beta_{2c-1}^{s,ols}, \beta_{2c}^{s,ols})$:

$$\beta_{2c-1}^{s,ols} = \vec{s}'_{2c-1}\vec{\alpha}^{s,c,ols} \text{ and } \beta_{2c}^{s,ols} = \vec{s}'_{2c}\vec{\alpha}^{s,c,ols} \tag{39}$$

While:

$$\beta_k^{s,ols} = \vec{s}'_k\vec{\alpha}^{s,k^*,ols} \tag{40}$$

We note that this algorithm yields the same results as reiterated solution of P_2^c, a claim we formalize as:

Lemma 7: Given $\left\{\beta_j^{s,ols}\right\}_{j\leq 2c-2}, \left\{\beta_j^{s-1,ols}\right\}_{2c-1<j\leq k}\Big\}$, and $\left\{y_i, \{x_{ij}\}_{j=1}^k\right\}_{i=1}^n$, the left hand side values of (39) and (40) correspond to solutions for $P2_c^s$ and $P3_c^s$ respectively.

Proof: See the appendix.

Notice that as we move toward a solution the $\vec{\alpha}^{s,c,ols}$ change, but S remains the same. For large values of n this can represent a substantial computational saving. Hastie[3] describes using a similar procedure, which he calls "covariance updating', for glmnet. Indeed, we suspect that, adjusting for differences in notation, the row of S that deals with the odd singleton variable in our framework coincides exactly with the algorithmic artfulness described by Hastie.

Managing the Active Set

Another important source of computational speed is the management of the "active set" used in the estimation. The idea is to restrict our attention to only variables that have a chance of surviving the LASSO process, and for this we have a straightforward screening procedure.

Firstly we identify the explanator max for which $r_{max,y} \geq r_{j,y} \forall j$, this is our starting value λ_{max} for λ. Our "active set" of variables consists solely of $\{x_{max,i}\}_{i=1}^n$. At $\lambda = r_{max,y}$ the LASSO with x_{max} as our sole potential explanator will produce a coefficient of zero, just barely censoring x_{max}. We then reduce λ by successive increments.

Now suppose that corresponding to the current value of λ we have an active set A_λ of variables $\{x_k\}_{k\in A_\lambda}$, we have estimated the LASSO coefficients corresponding to λ, and we are about to move

[3]See `http://web.stanford.edu/ hastie/TALKS/glmnet.pdf`

on to the next lower value, λ', in our sequence. Before we move on, for each variable x_j that is excluded from the active set we calculate:

$$b_j^{\lambda^{shadow}} = r_{y,j} - \sum_{k \in A_\lambda} r_{j,k} \hat{\beta}_k^{LASSO}(\lambda)$$

Next, we add any variable x_{j^*} for which $b_j^{\lambda^{shadow}} > \lambda'$ to the active list $A_{\lambda'}$ corresponding to λ'.

This simple rule screens out any variable that would not earn a positive LASSO coefficient at the next iteration if it was evaluated first using unicoordinate descent. This guarantees that there are always more observations than variables in the active set, and it keeps the remaining calculations to a minimum. We found that when we added this step it resulted in a nearly three fold acceleration of our algorithm.

Analytical inflection points and the Active Set

All of the preceding computational procedures are easily adapted to cases in which some of the variables are exempted from the LASSO, as might arise when one knows that a certain list of variables from a "reference model" need to be included in the specification. However, when all the variables are subject to the LASSO, we have one more computational arrow in our quiver–we can solve for the first two inflection points after λ_{max} at very low computation cost, bringing analytical formulas to bear. This enables us to jump quickly through the first segment of the LASSO trellis, providing another substantial boost to the speed of our algorithm.

The First Jump

On the interval between λ^{max} and the smallest λ value, $\lambda_{sidekick}$, that leaves but one nonzero LASSO coefficient we know that the coefficient for the nonzero LASSO coefficient is a linear function of λ:

$$\beta_{max} = sign(r_{max,y})\left(|r_{max,y}| - \tfrac{\lambda}{2}\right) = a_{max} + c_{max}\lambda$$

where $a_{max} = r_{max,y}$ and $c_{max} = -\frac{1}{2}sign(r_{max,y})$.

Over the same interval, the remaining OLS coefficients, conditional on β_{max}, are themselves linear in β_{max}:

$$\beta_j = r_{j,y} - r_{max,j}\beta_{max}$$

and hence they are also linear in λ:

$$\beta_j = (r_{j,y} - r_{max,j}a_{max}) + r_{max,j}c_{max}\lambda = a_j + c_j\lambda$$

where $a_j = r_{j,y} - r_{max,j}r_{max,y}$ and $c_j = -\frac{1}{2}sign(r_{max,y})r_{max,j}$.

Every variable except x_{max} will satisfy the following condition for $\lambda \in (\lambda_{sidekick}, \lambda_{max})$:

$$-\lambda < a_j + c_j\lambda < \lambda$$

Now let:

$$\lambda_j^- = \frac{a_j}{1 - c_j} \text{ and } \lambda_j^+ = \frac{-a_j}{1 + c_j}$$

while:

$$\lambda_j^* = \begin{cases} \lambda_j^- & \text{if } max\{0, \lambda_j^+\} < \lambda_j^- \leq \lambda_{max} \\ \lambda_j^+ & \text{if } max\{0, \lambda_j^-\} \leq \lambda_j^+ \leq \lambda_{max} \\ 0 & \text{otherwise} \end{cases}$$

It follows that:

$$\lambda_{sidekick} = max_j\{\lambda_j^*\}_{j \neq max}$$

If we let $x_{sidekick}$ denote the variable associated with this maximum value we see that at $\lambda_{sidekick}$ we have:

$$\beta_{max}^{LASSO} = a_{max} + c_{max}\lambda_{sidekick}$$

while all the other beta values are equal to zero. Notice that $A^{lambda_{sidekick}} = \{sidekick, max\}$.

The Second Jump

Now let's consider what happens for $\lambda \in (\lambda_{next}, \lambda_{sidekick})$, where λ_{next} corresponds to the next inflection point after $\lambda_{sidekick}$. Let $\{\hat{\alpha}_{max}, \hat{\alpha}_{sidekick}\}$ denote the coefficients from an OLS regression of y on x_{max} and $x_{sidekick}$. Along this interval we will have an interior solution for the LASSO coefficients corresponding to x_{max} and $x_{sidekick}$, which will thus be linear functions of λ:

$$\alpha_{max}^{LASSO} = \hat{\alpha}_{max} - sign(\hat{\alpha}_{max})\frac{\lambda}{2(1 + r_{max,sidekick})} \text{ and } \alpha_{sidekick}^{LASSO} = \hat{\alpha}_{sidekick} - sign(\hat{\alpha}_{sidekick})\frac{\lambda}{2(1 + r_{max,sidekick})}$$

while the conditional least squares estimator for each of the remaining coefficients is linear in α_{max}^{LASSO} and $\alpha_{sidekick}^{LASSO}$:

$$\hat{\beta}_j = r_{yj} - \alpha_{max}^{LASSO} r_{j,max} - \alpha_{sidekick}^{LASSO} r_{j,sidekick}$$

Substituting from our expressions for the two active coefficients this becomes:

$$\hat{\beta}_j = \tilde{a}_j + \tilde{c}_j \lambda$$

where:

$$\tilde{a}_j = r_{yj} - \hat{\alpha}_{max} r_{j,max} - \hat{\alpha}_{sidekick} r_{j,sidekick}$$

and:

$$\tilde{c}_j = \frac{sign(\hat{\alpha}_{max}) r_{j,max} + sign(\hat{\alpha}_{sidekick}) r_{j,sidekick}}{2(1 + r_{max,sidekick})}$$

We now proceed in parallel with the first update, every variable except x_{max} and $x_{sidekick}$ will satisfy the following condition for $\lambda \in (\lambda_{next}, \lambda_{sidekick})$:

$$-\lambda < \tilde{a}_j + \tilde{c}_j \lambda < \lambda$$

Now let:

$$\tilde{\lambda}_j^- = \frac{\tilde{a}_j}{1 - \tilde{c}_j} \text{ and } \tilde{\lambda}_j^+ = \frac{-\tilde{a}_j}{1 + \tilde{c}_j}$$

while:

$$\tilde{\lambda}_j^* = \begin{cases} \tilde{\lambda}_j^- & \text{if } max\{0, \tilde{\lambda}_j^+\} < \tilde{\lambda}_j^- \leq \tilde{\lambda}_{max} \\ \tilde{\lambda}_j^+ & \text{if } max\{0, \tilde{\lambda}_j^-\} \leq \tilde{\lambda}_j^+ \leq \tilde{\lambda}_{max} \\ 0 & \text{otherwise} \end{cases}$$

It follows that:

$$\lambda_{next} = max_j \{\tilde{\lambda}_j^*\}_{j \notin \{max, sidekick\}}$$

If we let x_{next} denote the variable associated with this maximum value we see that at λ_{next} we have:

$$\alpha_{max}^{LASSO} = \hat{\alpha}_{max} - sign(\hat{\alpha}_{max}) \frac{\lambda_{next}}{2(1 + r_{max,sidekick})} \text{ and } \alpha_{sidekick}^{LASSO} = \hat{\alpha}_{sidekick} - sign(\hat{\alpha}_{sidekick}) \frac{\lambda_{next}}{2(1 + r_{max,sidekick})}$$

while all the other beta values are equal to zero[4]. Notice that $A^{lambda_{next}} = \{next, sidekick, max\}$.

Comparative Timing

Our estimator is still at the developmental stage, and all of our code is written in R, whereas it's `glmnet` competitor is written in optimized FORTRAN. We expect the usual computational acceleration to emerge when we "translate" our code into C^{++}. However, we do want to provide the reader with a clearer understanding of the practical advantages of bicoordinate descent, and so we benchmark our estimator against unicoordinate descent using a variety of datasets that vary in size, and in the severity of the collinearity observed among their component variables.

[4]In the very unlikely event that $\lambda_{next} < \lambda_{drop} = 2sign(\hat{\alpha}_{max})(1 + r_{max,sidekick})\hat{\alpha}_{max}$ we instead stop at λ_{drop}, at which point the "max" variable goes dormant, and we repeat the second jump using sidekick in place of max, and λ_{drop} instead of $\lambda_{sidekick}$.

Speed Comparisons

	Data Passes	Min.	First Quintile	Mean	Median	Third Quintile	Max.
RED							
Bicoord.	121	15.82811	17.01458	18.62527	17.43571	18.32009	45.96997
Unicoord.	342	60.78047	62.07158	65.92969	62.99484	66.20146	125.62298
SOIL							
Bicoord.	245	43.91919	45.29539	46.01222	45.85828	46.47594	49.01433
Unicoord.	647	138.48964	143.93945	146.69634	144.66505	145.93290	173.56228
WHITE							
Bicoord.	253	51.66146	52.67688	53.69331	53.02394	53.33742	82.76253
Unicoord.	520	100.26352	102.47824	104.02415	102.89359	103.92719	131.41892

Unit: milliseconds Trials: 100

RED Wine quality data from Cortez et al. (2009)
`https://archive.ics.uci.edu/ml/datasets/Wine+Quality`

SOIL Soil Quality data from Bondell and Smith (2008) `http://www.biometrics.tibs.org`

WHITE Wine quality data from Cortez et al. (2009)
`https://archive.ics.uci.edu/ml/datasets/Wine+Quality`

Results of several time trials appear in the accompanying table. We used the `microbenchmark` package in R for the speed tests, with 100 trials. Estimation times for bicoordinate and unicoordinate descent are given in microseconds. We employ our warm starts for both unicoordinate and for bicoordinate descent, whereas we only apply the LARS updates and the management of the active set for bicoordinate descent.

The first column counts complete "passes" through the data, with each pass corresponding to a full set of parameter updates. Notice that this count is unaffected by the fact that bicoordinate descent updates parameters two at a time. If we have twenty parameters in our model, one data pass by bicoordinate descent consists of updating each of the ten pairs of parameters, whereas one data pass for unicoordinate descent involves twenty single parameter updates, either way twenty parameters are updated, and either way we count but a single pass through the data. Managing the active set also leaves our accounting for iterations unaffected (though it reduces the number of data passes we need to take), if we have twenty parameters with six active and fourteen dormant, then one round of updates to the six active parameters counts as a "data pass".

We observe a dramatic reduction in data passes moving from unicoordinate descent to bicoordinate descent, and a comparable reduction in the time required to conduct the calculations, with the bicoordinate algorithm working between two and three and a half times as fast. We note that this speed advantage comes despite the extra "overhead" costs of bicoordinate descent, which recalibrates the parameter matches every time a new parameter enters the active set.

3 Discussion

Given the advantages offered by exploitation of the correlations among the explanators, why should one stop at bicoordinate descent? Why not coordinate across even more variables at each step? Indeed, when the design matrix is of full rank the standard formula for calculating regression coefficients converges in but a single step. However, with a large number of explanators the constraint set of the LASSO becomes a high dimensional polytope with myriad corners, edges, and faces to check for possible solutions. Also, of course, the matrix inversion problem can be computationally intense when the design matrix is of full rank but large, while it becomes impossible when the matrix is nonsingular, as it is guaranteed to be for a sufficiently large number of explanators.

The huge appeal of one at a time coordinate wise descent is its robustness to the rank of the design matrix. Tibshirani's soft thresholding vastly streamlines the updating process, and it relies on the convenient result that the signs of the LASSO coefficient updates will never be opposite those of the signs of the unconstrained coordinate wise regression update steps.

The analogy to this "no sign reversal" condition in our formulation is that our pairwise LASSO updates are guaranteed to remain in the closure of the same quadrant as the pairwise regression coefficient updates. The cost of moving to bicoordinate descent is that it will only work for pairs of explanatory variables that are not perfectly correlated. But this is a scant price to pay, as the analyst has a variety of options; one of the perfectly correlated pair of variables could simply be dropped from the specification, or one could simply rematch the perfectly correlated pairs with other variables, or one could apply ordinary coordinate wise descent to the offending pairs.

Could this approach be extended to encompass tricoordinate descent? Perhaps, but the very convenient result that the LASSO updates will always be found in the same quadrants as the unconstrained updates does not generalize. In his figure 3a, Tibshirani (1996) p.271 shows that with three variables the LASSO coefficients may constitute interior solutions in a different quadrant than the regression coefficients. An interesting subject for ongoing research is to identify whether there are conditions on the correlations among triples of variables that guarantee that the LASSO updates

will be contained in the same octant as the least squares coefficient update steps.

Our currently active research extends the results in this paper to the case of weighted least squares. The case of weighted least squares is more challenging, because when the variances of the paired explanatory variables differ it is possible for the conditional LASSOed estimates to "escape" into the second or fourth quadrants. However, we are extending the results here to encompass those cases in a computationally efficient manner. The extension to weighted least squares is vital in part because we can formulate LASSOed logit and probit estimates as cases of iterated weighted least squares.

Appendix

Proof of Lemma 1

Proof.

$$
\begin{aligned}
& RSS_c\left(\beta_{2c-1}, \beta_{2c} | \left\{v_{ic}^s, x_{i,2c-1}, x_{i,2c}\right\}_{i=1}^n\right) \\
= & \sum_{i=1}^n \left(v_{ic}^s - \beta_{2c-1} x_{i,2c-1} - \beta_{2c} x_{i,2c}\right)^2 \\
= & \sum_{i=1}^n \left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c} - (\beta_{2c-1} - \beta_{2c-1}^{s,ols}) x_{i,2c-1} - (\beta_{2c} - \beta_{2c}^{s,ols}) x_{i,2c}\right)^2 \\
= & \sum_{i=1}^n \Bigg(\left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c}\right)^2 \\
& \quad + 2\left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c}\right)\left((\beta_{2c-1} - \beta_{2c-1}^{s,ols}) x_{i,2c-1} + (\beta_{2c} - \beta_{2c}^{s,ols}) x_{i,2c}\right) \\
& \quad - \left((\beta_{2c-1} - \beta_{2c-1}^{s,ols}) x_{i,2c-1} + (\beta_{2c} - \beta_{2c}^{s,ols}) x_{i,2c}\right)^2 \Bigg) \\
= & \sum_{i=1}^n \left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c}\right)^2 \\
& \quad + 2(\beta_{2c-1} - \beta_{2c-1}^{s,ols}) \left(\sum_{i=1}^n \left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c}\right) x_{i,2c-1}\right) \\
& \quad + 2(\beta_{2c} - \beta_{2c}^{s,ols}) \left(\sum_{i=1}^n \left(v_{ic}^s - \beta_{2c-1}^{s,ols} x_{i,2c-1} - \beta_{2c}^{s,ols} x_{i,2c}\right) x_{i,2c}\right) \\
& \quad + \sum_{i=1}^n \left((\beta_{2c-1} - \beta_{2c-1}^{s,ols}) x_{i,2c-1} + (\beta_{2c} - \beta_{2c}^{s,ols}) x_{i,2c}\right)^2
\end{aligned}
$$

but the least squares estimates are chosen to guarantee that:

$$\sum_{i=1}^{n}\left(v_{ic}^{s}-\beta_{2c-1}^{s,ols}x_{i,2c-1}-\beta_{2c}^{s,ols}x_{i,2c}\right)x_{i,2c-1}=0 \qquad \text{and} \qquad \sum_{i=1}^{n}\left(v_{ic}^{s}-\beta_{2c-1}^{s,ols}x_{i,2c-1}-\beta_{2c}^{s,ols}x_{i,2c}\right)x_{i,2c}=0$$

so our expression simplifies to:

$$\begin{aligned} & RSS_c\left(\beta_{2c-1},\beta_{2c}|\left\{v_{ic}^{s},x_{i,2c-1},x_{i,2c}\right\}_{i=1}^{n}\right) \\ = & \sum_{i=1}^{n}\left(v_{ic}^{s}-\beta_{2c-1}^{s,ols}x_{i,2c-1}-\beta_{2c}^{s,ols}x_{i,2c}\right)^2+\sum_{i=1}^{n}\left((\beta_{2c-1}-\beta_{2c-1}^{s,ols})x_{i,2c-1}+(\beta_{2c}-\beta_{2c}^{s,ols})x_{i,2c}\right)^2 \\ = & sse_0^{c,s}+Q(\beta_{2c-1}-\beta_{2c-1}^{s,ols},\beta_{2c}-\beta_{2c}^{s,ols},R_c) \end{aligned}$$

□

Proof of Lemma 2

Proof. When the constraint is not binding, the result is trivial and the OLS and LASSO estimates coincide, whereas if $\theta_c^s=0$ then the result again holds trivially, as the LASSO estimates must both equal zero. Now consider what happens when $0<\theta_c^s<|\beta_{2c-1}^{s,ols}|+|\beta_{2c}^{s,ols}|$. Any pair $\vec{z}_0=(z_{2c-1},z_{2c})$ such that $z_{2c-1}+z_{2c}=\alpha<\theta_c^s$ is dominated by $\vec{z}'=(z_{2c-1}+\frac{\theta_c^s-\alpha}{2},z_{2c}+\frac{\theta_c^s-\alpha}{2})$:

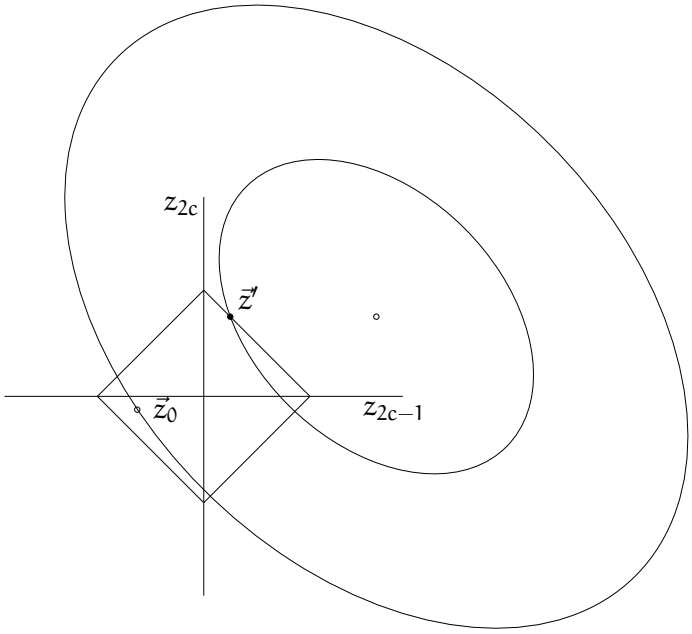

Figure 5: The unit simplex dominates the constraint set.

$$
\begin{aligned}
& Q\big(z_{2c-1} + \tfrac{\theta_c^s-\alpha}{2} - |\beta_{2c-1}^{s,ols}|, z_{2c} + \tfrac{\theta_c^s-\alpha}{2} - |\beta_{2c}^{s,ols}|, R_c^{s*}\big) - Q\big(z_{2c-1} - |\beta_{2c-1}^{s,ols}|, z_{2c} - |\beta_{2c}^{s,ols}|, R_c^{s*}\big) \\
= \;& \big(\tfrac{\theta_c^s-\alpha}{2}, \tfrac{\theta_c^s-\alpha}{2}\big) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} \frac{\theta_c^s-\alpha}{2} \\ \frac{\theta_c^s-\alpha}{2} \end{pmatrix} + 2\big(\tfrac{\theta_c^s-\alpha}{2}, \tfrac{\theta_c^s-\alpha}{2}\big) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} z_{2c-1} - \beta_1^{s,ols} \\ z_{2c} - \beta_2^{s,ols} \end{pmatrix} \\
= \;& 2(1 + R_c^{s*})\big(\tfrac{\theta_c^s-\alpha}{2}\big)^2 + 2(1 + R_c^{s*})\big(\tfrac{\theta_c^s-\alpha}{2}\big)\big\{z_{2c-1} + z_{2c} - \beta_1^{s,ols} - \beta_2^{s,ols}\big\} \\
= \;& (1 + R_c^{s*})(\theta_c^s - \alpha)\big\{\tfrac{\theta_c^s-\alpha}{2} + \alpha - \theta_c^s\big\} \\
= \;& -(1 + R_c^{s*})\big\{\frac{(\theta_c^s - \alpha)^2}{2}\big\} \\
< \;& 0
\end{aligned}
$$

Thus the only portion of the constraint that is not dominated according to this argument is the line segment S:

$$S = \{(z_1, z_2) | z_1 \geq 0, z_2 \geq 0, z_1 + z_2 = \theta_c^s\} \tag{41}$$

hence the solution to PZ: $(\hat{z}_{2c-1}, \hat{z}_{2c}) \in S$.

Finally we need to check that if $z_{2c-1} + z_{2c} = \alpha < \theta_c^s$ is inside the constraint set, then so is $(z_{2c-1} + \frac{\theta_c^s-\alpha}{2}, z_{2c} + \frac{\theta_c^s-\alpha}{2})$. We know that the constraint $|z_1| + |z_2| \leq \theta_c^s$ can be rewritten in terms of the two conditions: $C1 : -\theta_c^s \leq z_1 + z_2 \leq \theta_c^s$ and $C2 : -\theta_c^s \leq z_1 - z_2 \leq \theta_c^s$. The pair $(z_{2c-1} + \frac{\theta_c^s-\alpha}{2}, z_{2c} + \frac{\theta_c^s-\alpha}{2})$ satisfy C1 by construction, while $z_{2c-1} + \frac{\theta_c^s-\alpha}{2} - (z_{2c} + \frac{\theta_c^s-\alpha}{2}) = z_{2c-1} - z_{2c}$ so that if $-\theta_c^s \leq z_{2c-1} - z_{2c} \leq \theta_c^s$ it follows that $\theta_c^s \leq z_{2c-1} + \frac{\theta_c^s-\alpha}{2} - (z_{2c} + \frac{\theta_c^s-\alpha}{2}) \leq \theta_c^s$

□

Proof of Lemma 3

Proof. Differentiating our expression for Q, (15), we have:

$$\frac{\partial Q}{\partial z_{2c-1}} = 2(z_{2c-1} - |\beta_{2c-1}|) + 2R_c^{s*}(z_{2c} - |\beta_{2c}|) \quad \text{and} \quad \frac{\partial Q}{\partial z_2} = 2(z_{2c} - |\beta_{2c}|) + 2R_c^{s*}(z_{2c-1} - |\beta_{2c-1}|)$$

substituting into (16) this yields:

$$
\begin{aligned}
\frac{dz_{2c}}{dz_{2c-1}} &= -\frac{2(z_{2c-1} - |\beta_{2c-1}|) + 2R_c^{s*}(z_{2c} - |\beta_{2c}|)}{2(z_{2c} - |\beta_{2c}|) + 2R_c^{s*}(z_{2c-1} - |\beta_{2c-1}|)} \\
&= -\frac{(z_{2c-1} - |\beta_{2c-1}|) + R_c^{s*}(z_{2c} - |\beta_{2c}|)}{(z_{2c} - |\beta_{2c}|) + R_c^{s*}(z_{2c-1} - |\beta_{2c-1}|)}
\end{aligned}
$$

□

Proof of Lemma 4

Proof. Substituting from 12) we have:

$$\begin{aligned} & Q(\theta_c^s - |\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|, R_c^*) \\ = \; & \left(\theta_c^s - |\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|\right) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} \theta_c^s - |\beta_{2c-1}^{s,ols}| \\ -|\beta_{2c}^{s,ols}| \end{pmatrix} \\ = \; & -2\theta_c^s(|\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}|) + \left(\theta_c^{s2} + \left(-|\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|\right) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} -|\beta_{2c-1}^{s,ols}| \\ -|\beta_{2c}^{s,ols}| \end{pmatrix}\right) \quad (42) \end{aligned}$$

likewise:

$$\begin{aligned} & Q(-|\beta_{2c-1}^{s,ols}|, \theta_c^s - |\beta_{2c}^{s,ols}|, R_c^*)) \\ = \; & \left(-|\beta_{2c-1}^{s,ols}|, \theta_c^s - |\beta_{2c}^{s,ols}|\right) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} -|\beta_{2c-1}^{s,ols}| \\ \theta_c^s - |\beta_{2c}^{s,ols}| \end{pmatrix} \\ = \; & -2\theta_c^s(R_c^{s*}|\beta_{2c-1}^{s,ols}| + |\beta_{2c}^{s,ols}|) + \left(\theta_c^{s2} + \left(-|\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|\right) \begin{pmatrix} 1 & R_c^{s*} \\ R_c^{s*} & 1 \end{pmatrix} \begin{pmatrix} -|\beta_{2c-1}^{s,ols}| \\ -|\beta_{2c}^{s,ols}| \end{pmatrix}\right) \quad (43) \end{aligned}$$

Calculating the difference between (42) and (43) we have:

$$\begin{aligned} & Q(\theta_c^s - |\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|, R_c^*) - Q(-|\beta_{2c-1}^{s,ols}|, \theta_c^s - |\beta_{2c}^{s,ols}|, R_c^*)) \\ = \; & 2\theta_c^s\Big(|\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}| - \left(R_c^{s*}|\beta_{2c-1}^{s,ols}| + |\beta_{2c}^{s,ols}|\right)\Big) \\ = \; & n2\theta_c^s(1 - R_c^{s*})\left(|\beta_{2c-1}^{s,ols}| - |\beta_{2c}^{s,ols}|\right) \end{aligned}$$

However, $|R_c^{s*}| < 1$, recall that our data contain no perfectly correlated pairs. Likewise $\theta_c^s > 0$ by assumption, and so $2\theta_c^s(1 - R_c^{s*} > 0$, hence we have:

$$\begin{aligned} & sign\Big(Q(\theta_c^s - |\beta_{2c-1}^{s,ols}|, -|\beta_{2c}^{s,ols}|, R_c^*) - Q(-|\beta_{2c-1}^{s,ols}|, \theta_c^s - |\beta_{2c}^{s,ols}|, R_c^*))\Big) \\ = \; & sign\left(|\beta_{2c-1}^{s,ols}| - |\beta_{2c}^{s,ols}|\right) \end{aligned}$$

□

Proof of Lemma 5

Proof. Considering cases for which $\lambda > 0$, at an interior solution the non-negativity constraints are not binding, so that $\mu_{2c-1} = \mu_{2c} = 0$. Differentiating (26) with respect to z_{2c-1}, z_{2c}, and λ we have:

$$\begin{aligned}
\frac{\partial L}{\partial z_{2c-1}} &= 2(z_{2c-1} - |\beta_{2c-1}^{s,ols}|) + 2R_c(z_{2c} - |\beta_{2c-1}^{s,ols}|) + \lambda = 0 \\
\frac{\partial L}{\partial z_{2c}} &= 2R_c(z_{2c-1} - |\beta_{2c-1}^{s,ols}|) + 2(z_{2c} - |\beta_{2c-1}^{s,ols}|) + \lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= z_{2c-1} + z_{2c} - \theta_c^s = 0
\end{aligned} \tag{44}$$

If we add the first two equations:

$$2(1 + R_c)(z_{2c-1} - |\beta_{2c-1}^{s,ols}|) + 2(1 + R_c)(z_{2c} - |\beta_{2c}^{s,ols}|) + 2\lambda = 0$$

rearranging terms this becomes:

$$2(1 + R_c)(z_{2c-1} + z_{2c}) - 2(1 + R_c)|\beta_{2c-1}^{s,ols}| - 2(1 + R_c)|\beta_{2c}^{s,ols}| + 2\lambda = 0 \tag{45}$$

now substitute from $\frac{\partial L}{\partial \lambda} = 0$ to obtain:

$$2(1 + R_c)\theta_c^s - 2(1 + R_c)|\beta_{2c-1}^{s,ols}| - 2(1 + R_c)|\beta_{2c}^{s,ols}| + 2\lambda = 0 \tag{46}$$

solving for θ_c^s we have:

$$\theta_c^s = |\beta_{2c-1}^{s,ols}| + |\beta_{2c}^{s,ols}| - \frac{\lambda}{1 + R_c}$$

□

Proof of Lemma 6

Proof. Turning to our first order conditions for (26) we require:

$$\begin{aligned}
\frac{\partial L}{\partial z_{2c-1}} &= 2(z_{2c-1} - |\beta_{2c-1}^{s,ols}|) + 2R_c(0 - |\beta_{2c}^{s,ols}|) + \lambda = 0 \\
\frac{\partial L}{\partial z_{2c}} &= 2R_c(z_{2c-1} - |\beta_{2c-1}^{s,ols}|) + 2(0 - |\beta_{2c}^{s,ols}|) + \lambda \geq 0 \\
\frac{\partial L}{\partial \lambda} &= z_{2c-1} + 0 - \theta_c^s = 0
\end{aligned} \tag{47}$$

Substituting z_{2c-1} from the third expression into the first and solving for θ_c^s yields:

$$\theta_c^s = |\beta_{2c-1}^{s,ols}| + R_c^{s*}|\beta_{2c}^{s,ols}| - \frac{\lambda}{2}$$

□

Proof of Lemma 7

Proof. Start with the update for $(\beta^{s,ols}_{2c-1}, \beta^{s,ols}_{2c})$. We solve:

$$\min_{\beta^*_{2c-1},\beta^*_{2c}} \sum_{i=1}^{n} \big(y_i - \sum_{j\notin\{2c-1,2c\}} x_{j,i}\beta^0_j - x_{2c-1,i}\beta^*_{2c-1} - x_{2c,i}\beta^*_{2c}\big)^2$$

where $\beta^0_j = \beta^{s,ols}_j$ if $j < 2c-1$, while for $2c < j$ we let $\beta^0_j = \beta^{s-1,ols}_j$.

This leads to the following first order conditions:

$$\begin{aligned} -2\sum_{i=1}^{n} \big(y_i - \sum_{j\notin\{2c-1,2c\}} x_{j,i}\beta^0_j - x_{2c-1,i}\beta^*_{2c-1} - x_{2c,i}\beta^*_{2c}\big)x_{2c-1,i} &= 0 \\ -2\sum_{i=1}^{n} \big(y_i - \sum_{j\notin\{2c-1,2c\}} x_{j,i}\beta^0_j - x_{2c-1,i}\beta^*_{2c-1} - x_{2c,i}\beta^*_{2c}\big)x_{2c,i} &= 0 \end{aligned}$$

Using our newly developed notation we can rewrite these conditions as:

$$\begin{aligned} -2\big(r_{y,2c-1} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c-1}\beta^0_j - \beta^*_{2c-1} - r_{2c-1,2c}\beta^*_{2c}\big) &= 0 \\ -2\big(r_{y,2c} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c}\beta^0_j - r_{2c-1,2c}\beta^*_{2c-1} - \beta^*_{2c}\big) &= 0 \end{aligned}$$

That is:

$$\begin{pmatrix} 1 & r_{2c-1,2c} \\ r_{2c-1,2c} & 1 \end{pmatrix} \begin{pmatrix} \beta^*_{2c-1} \\ \beta^*_{2c} \end{pmatrix} = \begin{pmatrix} r_{y,2c-1} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c-1}\beta^0_j \\ r_{y,2c} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c}\beta^0_j \end{pmatrix}$$

This simplifies to:

$$\begin{pmatrix} \beta^*_{2c-1} \\ \beta^*_{2c} \end{pmatrix} = \frac{1}{1-r^2_{2c,2c-1}} \begin{pmatrix} 1 & -r_{2c-1,2c} \\ -r_{2c-1,2c} & 1 \end{pmatrix}^{-1} \begin{pmatrix} r_{y,2c-1} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c-1}\beta^0_j \\ r_{y,2c} - \sum_{j\notin\{2c-1,2c\}} r_{j,2c}\beta^0_j \end{pmatrix}$$

That is:

$$\begin{aligned} \begin{pmatrix} \beta^*_{2c-1} \\ \beta^*_{2c} \end{pmatrix} &= \frac{1}{1-r^2_{2c,2c-1}} \begin{pmatrix} (r_{y,2c-1} - r_{2c,2c-1}r_{y,2c}) - \sum_{j\notin\{2c-1,2c\}} (r_{j,2c-1} - r_{j,2c}r_{2c-1,2c})\beta^0_j \\ (r_{y,2c} - r_{2c,2c-1}r_{y,2c-1}) - \sum_{j\notin\{2c-1,2c\}} (r_{j,2c} - r_{j,2c}r_{2c-1,2c})\beta^0_j \end{pmatrix} \\ &= \begin{pmatrix} \vec{s}'_{2c-1}\vec{\alpha}^{s,c,ols} \\ \vec{s}'_{2c}\vec{\alpha}^{s,c,ols} \end{pmatrix} \end{aligned} \tag{48}$$

But this is simply (39). Similar, and even more straightforward calculations show that (40) corresponds to the solution for $\mathsf{P3}_{\mathsf{c}}^{\mathsf{s}}$.

□

References

Bondell, Howard D. and Brian J. Smith. 2008. "Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR." Biometrics 64:115–23.

Cortez, P., A. Cerdeira, F. Almeida, T. Matos and J. Reis. 2009. "Modeling wine preferences by data mining from physicochemical properties." Decision Support Systems 47:547–553.

Friedman, Jerome H., Trevor Hastie and Rob Tibshirani. 2010a. "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software 33:1–22.

Friedman, Jerome, Trevor Hastie and Robert Tibshirani. 2010b. "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software 33:1–22.
URL: *http://www.jstatsoft.org/v33/i01/*

Fu, Wenjiang. 1998. "Penalized Regressions: The Bridge vs the Lasso." Journal of Computational and Graphical Statistics 7:397–416.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society, Series B. 58:267–88.