SUPPLEMENT TO "A LASSO FOR HIERARCHICAL INTERACTIONS"

By Jacob Bien*, Jonathan Taylor† and Robert Tibshirani‡

Stanford University

1. Effect of constraint.

For notational simplicity, we write $r(\beta^+, \beta^-, \Theta) \in \mathbb{R}^n$ to denote the residuals, $y - \hat{y}(\beta^+, \beta^-, \Theta)$, as a function of the parameters. The strong Hierarchical Lasso problem is the following:

$$\underset{\beta^+,\beta^-,\Theta}{\text{Minimize}} \quad \frac{1}{2}\|r(\beta^+, \beta^-, \Theta)\|^2 + \lambda_1 1^T(\beta^+ + \beta^-) + \lambda_2 \sum_j \|\Theta_j\|_1$$
$$\text{s.t.} \quad \|\Theta_j\|_1 \le \beta_j^+ + \beta_j^- \text{ and } \beta_j^+ \ge 0,\ \beta_j^- \ge 0 \text{ for each } j,\ \Theta = \Theta^T.$$

The Lagrangian is

$$\begin{aligned} L(\phi; \alpha, S, \gamma^\pm, U) &= \frac{1}{2}\|r(\beta^+, \beta^-, \Theta)\|^2 + \lambda_1 1^T(\beta^+ + \beta^-) + \lambda_2 \langle U, \Theta\rangle \\ &\quad + \sum_j \alpha_j (U_j^T \Theta_j - \beta_j^+ - \beta_j^-) - \gamma_j^+ \beta_j^+ - \gamma_j^- \beta_j^- + \langle S, \Theta - \Theta^T \rangle \\ &= \frac{1}{2}\|r(\beta^+, \beta^-, \Theta)\|^2 + (\lambda_1 1 - \alpha - \gamma^+)^T \beta^+ + (\lambda_1 1 - \alpha - \gamma^-)^T \beta^- \\ &\quad + \langle S - S^T + \operatorname{diag}(\lambda_2 1 + \alpha) U, \Theta \rangle, \end{aligned}$$

where $\alpha, \gamma^\pm, S, U$ are dual variables. According to the KKT conditions, $(\hat{\phi}; \hat{\alpha}, \widehat{S}, \hat{\gamma}^\pm, \widehat{U})$ is an optimal primal-dual pair if and only if

$$\begin{aligned} \pm x_j^T r(\hat\beta^+, \hat\beta^-, \widehat\Theta) &= \lambda_1 - \hat\alpha_j - \hat\gamma_j^\pm \\ (x_j * x_k)^T r(\hat\beta^+, \hat\beta^-, \widehat\Theta)/2 &= (\lambda_2 + \hat\alpha_j)\widehat{U}_{jk} + \widehat{S}_{jk} - \widehat{S}_{kj} \\ 0 = \hat\beta_j^\pm \hat\gamma_j^\pm \qquad 0 &= \hat\alpha_j(\|\widehat\Theta_j\|_1 - \hat\beta_j^+ - \hat\beta_j^-) \\ \widehat\Theta = \widehat\Theta^T, \qquad \hat\beta^\pm \ge 0, \qquad & \|\widehat\Theta_j\|_1 \le \hat\beta_j^+ + \hat\beta_j^- \qquad \hat\alpha, \hat\gamma^\pm \ge 0 \end{aligned}$$

$$\widehat{U}_{jk} = \begin{cases} \operatorname{sign}(\widehat\Theta_{jk}) & \widehat\Theta_{jk} \ne 0 \\ \in [-1, 1] & \widehat\Theta_{jk} = 0. \end{cases}$$

*Supported in part by the Gerald J. Lieberman Fellowship

†Supported by National Science Foundation Grant DMS-0906801 and AFOSR Grant 113039

‡Supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183

Now, letting $r^{(-j)} = r(\hat\beta^+, \hat\beta^-, \widehat\Theta) + (\hat\beta_j^+ - \hat\beta_j^-)x_j$ and recalling that $\|x_j\|^2 = 1$, there are three cases to consider:

1. $\hat\beta_j^+ \geq 0,\ \hat\beta_j^- = 0$:
$$x_j^T(r^{(-j)} - \hat\beta_j^+ x_j) = \lambda_1 - \hat\alpha_j - \hat\gamma_j^+ \implies \hat\beta_j^+ = [x_j^T r^{(-j)} - (\lambda_1 - \hat\alpha_j)]_+$$
Note that this case applies when $x_j^T r^{(-j)} \geq \lambda_1 - \hat\alpha_j$. Thus, in this case $\hat\beta_j^+ - \hat\beta_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat\alpha_j)$.
2. $\hat\beta_j^+ = 0,\ \hat\beta_j^- \geq 0$:
$$-x_j^T(r^{(-j)} + \hat\beta_j^- x_j) = \lambda_1 - \hat\alpha_j - \hat\gamma_j^- \implies \hat\beta_j^- = [-x_j^T r^{(-j)} - (\lambda_1 - \hat\alpha_j)]_+$$
Note that this case applies when $x_j^T r^{(-j)} \leq -(\lambda_1 - \hat\alpha_j)$. Thus, once again $\hat\beta_j^+ - \hat\beta_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat\alpha_j)$.
3. $\hat\beta_j^+ > 0,\ \hat\beta_j^- > 0 \quad (\implies \hat\gamma_j^+ = 0,\ \hat\gamma_j^- = 0)$
$$\pm x_j^T(r^{(-j)} - (\hat\beta_j^+ - \hat\beta_j^-)x_j) = \lambda_1 - \hat\alpha_j \implies \hat\beta_j^+ - \hat\beta_j^- = x_j^T r^{(-j)}.$$
Note that this case applies when $\hat\alpha_j = \lambda_1$, so trivially $\hat\beta_j^+ - \hat\beta_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat\alpha_j)$.

Thus, we have shown that $\hat\beta_j^+ - \hat\beta_j^- = \mathcal{S}(x_j^T r^{(-j)},\ \lambda_1 - \hat\alpha_j)$.

We can get rid of $\hat S$ by rewriting the subgradient equation involving it as
$$(x_j * x_k)^T r(\hat\beta^+, \hat\beta^-, \widehat\Theta) = (2\lambda_2 + \hat\alpha_j + \hat\alpha_k)\widehat U_{jk}$$
(note that symmetry in $\widehat\Theta$ implies that there exists a symmetric $\widehat U$).

Now, letting $r^{(-jk)} = r(\hat\beta^+, \hat\beta^-, \widehat\Theta) + (x_j * x_k)(\widehat\Theta_{jk} + \widehat\Theta_{kj})/2$, we get
$$\widehat\Theta_{jk}\|x_j * x_k\|^2 = (x_j * x_k)^T r^{(-jk)} - (2\lambda_2 + \hat\alpha_j + \hat\alpha_k)\widehat U_{jk} = \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2\lambda_2 + \hat\alpha_j + \hat\alpha_k).$$
This completes the proof for the Strong Hierarchical Lasso. Note that in the Weak Hierarchical Lasso case, the KKT conditions are identical except we do not have the constraint $\widehat\Theta = \widehat\Theta^T$ and we take $\widehat S = 0$. Thus, the relevant condition is simply
$$(x_j * x_k)^T r(\hat\beta^+, \hat\beta^-, \widehat\Theta) = 2(\lambda_2 + \hat\alpha_j)\widehat U_{jk} = 2(\lambda_2 + \hat\alpha_k)\widehat U_{kj}.$$
Note that the second equality implies that $\widehat U_{jk}\widehat U_{kj} \geq 0$ (since $\hat\alpha \geq 0$) and that if $|U_{jk}| = 1$, then $\hat\alpha_j \leq \hat\alpha_k$ and vice versa. Rearranging terms, we have
$$\begin{aligned}(\widehat\Theta_{jk} + \widehat\Theta_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat\alpha_j)\widehat U_{jk}\\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat\alpha_k)\widehat U_{kj}.\end{aligned}$$

Now, $\widehat{U}_{jk}\widehat{U}_{kj} \geq 0$ implies $\widehat{\Theta}_{jk}\widehat{\Theta}_{kj} \geq 0$ which implies that $(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})/2$, if nonzero, has the same sign as whichever of $\widehat{\Theta}_{jk}$ or $\widehat{\Theta}_{kj}$ (or both) is nonzero.

There are four cases:

1. $\widehat{\Theta}_{jk} \neq 0$, $\widehat{\Theta}_{kj} = 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_j \leq \hat{\alpha}_k$ since $|\widehat{U}_{jk}| = 1$.
2. $\widehat{\Theta}_{jk} = 0$, $\widehat{\Theta}_{kj} \neq 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_k) \cdot \mathrm{sign}(\widehat{\Theta}_{kj}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_k) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_k \leq \hat{\alpha}_j$ since $|\widehat{U}_{kj}| = 1$.
3. $\widehat{\Theta}_{jk} \neq 0$, $\widehat{\Theta}_{kj} \neq 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk}) \\ &= (x_j * x_k)^T r^{(-jk)} - 2(\lambda_2 + \hat{\alpha}_j) \cdot \mathrm{sign}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\end{aligned}$$
and $\hat{\alpha}_j = \hat{\alpha}_k$ since $|\widehat{U}_{jk}| = |\widehat{U}_{kj}| = 1$.
4. $\widehat{\Theta}_{jk} = 0$, $\widehat{\Theta}_{kj} = 0$:
$$\begin{aligned}(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 &= 0 \\ &= \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \hat{\alpha}_j)) \\ &= \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \hat{\alpha}_k))\end{aligned}$$
where the latter two equalities follow since $|(x_j * x_k)^T r^{(-jk)}| \leq 2(\lambda_2 + \hat{\alpha}_j)$ and $|(x_j * x_k)^T r^{(-jk)}| \leq 2(\lambda_2 + \hat{\alpha}_k)$.

We can encapsulate all of this into a single, simple expression:
$$(\widehat{\Theta}_{jk} + \widehat{\Theta}_{kj})\|x_j * x_k\|^2/2 = \mathcal{S}((x_j * x_k)^T r^{(-jk)},\ 2(\lambda_2 + \min\{\hat{\alpha}_j, \hat{\alpha}_k\})).$$

2. Proof that (5) **and** (6) **are equivalent.** We rewrite (5) in terms of $\beta = \beta^+ - \beta^-$ rather than β^-:
$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0, \beta, \Theta) + \lambda 1^T(2\beta^+ - \beta) + \frac{\lambda}{2}\|\Theta\|_1 \\ \text{s.t.} \quad & \Theta = \Theta^T,\ \beta^+ \geq 0,\ \beta^+ \geq \beta,\ \|\Theta_j\|_1 \leq 2\beta_j^+ - \beta_j\end{aligned}$$

or

$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta)+\lambda 1^T(2\beta^+-\beta)+\frac{\lambda}{2}\|\Theta\|_1\\ \text{s.t.} \quad & \Theta=\Theta^T,\ \max\{[\beta_j]_+,\ (\|\Theta_j\|_1+\beta_j)/2\}\le\beta_j^+,\end{aligned}$$

where $[\beta_j]_+ = \max\{\beta_j, 0\}$ is the positive part of β_j. This problem is the epigraph form of

$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta,\beta^+\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0,\beta,\Theta)+\lambda\sum_{j=1}^p(\max\{2[\beta_j]_+,\|\Theta_j\|_1+\beta_j\ \}-\beta_j)+\frac{\lambda}{2}\|\Theta\|_1\\ \text{s.t.} \quad & \Theta=\Theta^T\end{aligned}$$

which reduces to (6) since $2[\beta_j]_+ - \beta_j = |\beta_j|$.

3. Solving the logistic regression problem. For notational simplicity, in this section we use $\widetilde{X}$ and ϕ to denote the full data matrix and parameter combining both main effects and interactions. The binomial negative log-likelihood is

$$\ell(\beta_0,\phi)=-\sum_{i=1}^n[y_i\log p_i+(1-y_i)\log(1-p_i)]$$

where $p_i = p_i(\beta_0, \phi) = 1/(1 + e^{-\beta_0-\tilde{x}_i^T\phi})$. Now,

$$\frac{\partial\ell(\beta_0,\phi)}{\partial\beta_0}=-1^T(y-p) \qquad \nabla_\phi\ell(\beta_0,\phi)=-\widetilde{X}^T(y-p).$$

Thus, to solve $\min_{\beta_0,\phi}\ell(\beta_0, \phi) + h(\phi)$, we can use generalized gradient descent, which iteratively solves

$$\begin{pmatrix}\hat\beta_0^{(k)}\\ \hat\phi^{(k)}\end{pmatrix}=\arg\min_{\beta_0,\phi}\frac{1}{2t}\left\|\begin{pmatrix}\beta_0\\ \phi\end{pmatrix}-\left[\begin{pmatrix}\hat\beta_0^{(k-1)}\\ \hat\phi^{(k-1)}\end{pmatrix}+t\begin{pmatrix}1^T[y-p(\hat\beta_0^{(k-1)},\hat\phi^{(k-1)})]\\ \widetilde{X}^T[y-p(\hat\beta_0^{(k-1)},\hat\phi^{(k-1)})]\end{pmatrix}\right]\right\|^2+h(\phi).$$

This separates into two parts:

$$\begin{aligned}\hat\beta_0^{(k)}&=\hat\beta_0^{(k-1)}+t1^T[y-p(\hat\beta_0^{(k-1)},\hat\phi^{(k-1)})]\\ \hat\phi^{(k)}&=\text{Prox}_{2t\cdot h}\left(\hat\phi^{(k-1)}+t\widetilde{X}^T[y-p(\hat\beta_0^{(k-1)},\hat\phi^{(k-1)})]\right),\end{aligned}$$

where $\text{Prox}_{2t\cdot h}$ refers to the minimizer of (11). Looking at Algorithm 1, we see that this algorithm is identical except that for each k we update the estimate of the intercept and that we compute the residual as $y - p(\hat\beta_0, \hat\phi)$. The “difficult” part of the computation is identical!

4. ADMM for Strong Hierarchical Lasso. The ADMM algorithm has three parts:

1. Update $(\beta_0,\ \beta^{\pm},\ \Theta)$ by solving

$$\begin{aligned}\underset{\beta_0\in\mathbb{R},\ \beta^{\pm}\in\mathbb{R}^p,\ \Theta\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad & q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1 \\ & + \operatorname{tr}[U(\Theta - \widehat{\Omega})] + (\rho/2)\|\Theta - \widehat{\Omega}\|_F^2 \\ \text{s.t.} \quad & \beta_j^+ \geq 0, \beta_j^- \geq 0 \text{ for } j = 1, \ldots, p.\end{aligned}$$

   As with Algorithm 1, we may apply generalized gradient descent and `ONEROW` to solve this, but replacing the argument $\widetilde{\Theta}_j$ of `ONEROW` with $\delta\widehat{\Theta}_j^{(k-1)} - tZ_{(j,\cdot)}^T\hat{r}^{(k-1)} + \rho(\widehat{\Theta}_j^{(k-1)} - \widehat{\Omega}) + U$.
2. Update Ω by solving

$$\underset{\Omega\in\mathbb{R}^{p\times p}}{\text{Minimize}} \quad \operatorname{tr}[U(\widehat{\Theta} - \Omega)] + (\rho/2)\|\widehat{\Theta} - \Omega\|_F^2 \quad \text{s.t.} \quad \Omega = \Omega^T.$$

   This has the analytic solution $\widehat{\Omega} \leftarrow \frac{1}{2}(\widehat{\Theta} + \widehat{\Theta}^T) + \frac{1}{2\rho}(U + U^T)$.
3. Update $U \leftarrow U + \rho(\widehat{\Theta} - \widehat{\Omega})$:

Algorithm 2 in the paper gives the full algorithm.

Department of Biological Statistics
and Computational Biology
and Department of Statistical Science
Cornell University
Ithaca, NY 14853
E-mail: jbien@cornell.edu

Department of Statistics
Stanford University
Stanford, CA 94305
E-mail: jonathan.taylor@stanford.edu

Department of Health, Research, & Policy
and Department of Statistics
Stanford University
Stanford, CA 94305
E-mail: tibs@stanford.edu