

Supplementary material for Forward-LASSO with Adaptive Shrinkage

PETER RADCHENKO AND GARETH M. JAMES *

1 Proof of Claim 1

Let matrix Σ_{KK} , defined as $X_K^T X_K$, contain the correlations among the signal variables, and let $\mathbf{1}$ be a vector of 1's that has length S . Note that $\Sigma_{KK}\mathbf{1} = [1 - (S-1)\rho]\mathbf{1}$, which implies $\Sigma_{KK}^{-1}\mathbf{1} = [1 - (S-1)\rho]^{-1}\mathbf{1}$ and, consequently,

$$\nu_j(\rho) = X_j^T X_K \Sigma_{KK}^{-1} \text{sign}(\boldsymbol{\beta}) = \rho S [1 - (S-1)\rho]^{-1}, \quad \text{for } j \in K^c.$$

Note that $\nu_j([2S-1]^{-1}) = 1$ and $\nu_j(\rho)$ is a strictly increasing function for $\rho < [S-1]^{-1}$.

The proof of Theorem 2 in Wainwright (2009) establishes that for each value of the tuning parameter λ the necessary condition for the signed support recovery is $|\nu_j + \tilde{Z}_j| \leq 1$ with $\tilde{Z}_j = \lambda^{-1} X_j^T \Pi_{X_K^\perp}(\boldsymbol{\epsilon})$. Note that Z_j follows a non-degenerate zero-mean gaussian distribution for $S < n$, thus the probability of the above condition is at most 1/2. When $S \geq n$ the Lasso automatically fails, because none of its estimators can have more than $n - 1$ nonzero coefficients.

2 Proof of Theorem 1

Let K_1 and K_2 index the “large” and the “small” coefficients, respectively, and define $m_1 = \min\{|\beta_j|, j \in K_1\}$ and $M_2 = \max\{|\beta_j|, j \in K_2\}$. To substantially simplify the notation we will assume $q_1 + q_2 = 1$ from here on. The general case can be handled

*Marshall School of Business, University of Southern California. This work was partially supported by NSF Grant DMS-0705312.

with small modifications to the proof that follows. We can think of the block FLASH procedure as two applications of the Lasso, with the second application placing zero penalty on the variables selected at the first stage. For convenience, we will refer to the tuning parameters used to identify the two blocks of variables as $\lambda_1 = c_4 M_2$ and $\lambda_2 = c_2 \sqrt{\log p}$, respectively. We set $\lambda_2 = (1 + \epsilon)(2/\xi) \sqrt{2\sigma^2 \log p}$ for some positive constant ϵ . We will define λ_1 at the end of the proof. Inequality $\lambda_1 > \lambda_2$ will automatically follow from the definition.

We start with the following result that takes care of the error terms.

Lemma 1 *Let $W = \lambda_2^{-1} \|X_{K^c}^T \Pi_{X_K^\perp}(\epsilon)\|_\infty$ and $V = \|\Sigma_{KK}^{-1} X_K^T \epsilon\|_\infty$. Define \tilde{W} and \tilde{V} analogously but with K replaced by K_1 . Then the probability of the set*

$$\{W \vee \tilde{W} < \xi/2, V \vee \tilde{V} < \lambda_2 4\sqrt{2}\sigma\} \quad (7)$$

goes to one as p tends to infinity.

This lemma is a direct consequence of the corresponding results in Wainwright's Section III.C, using the fact that the minimal eigenvalues of Σ_{KK} and $\Sigma_{K_1 K_1}$ are bounded below by $S/(2S - 1)$ and $q_1 S/(2q_1 S - 1)$, respectively.

To complete the proof of Theorem 1 it is left to establish the next result.

Lemma 2 *The following statements are guaranteed to hold given (7).*

1. *Consider the weighted Lasso problem where the weights corresponding to K_1 are set to zero, and the weights corresponding to K_1^c are set to one. The solution for $\lambda = \lambda_2$ correctly recovers the signed support of β .*
2. *The nonzero coefficients of the Lasso solution corresponding to $\lambda = \lambda_1$ are indexed by K_1 .*

3 Proof of Lemma 2.1

The optimization problem of interest is

$$\arg \min_{\tilde{\beta}} \|Y - X\tilde{\beta}\|_2^2 + \lambda \sum_{k \in K_1^c} |\tilde{\beta}_k|. \quad (8)$$

Define vector \mathbf{s} by replacing the elements of $\text{sign}(\boldsymbol{\beta}_K)$ corresponding to the coefficients in K_1 with zeros. Let $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{h}$, where $\mathbf{h}_K = \Sigma_{KK}^{-1}(-\lambda\mathbf{s} + X_K^T\boldsymbol{\epsilon})$ and $\mathbf{h}_{K^c} = 0$. We will show that $\widehat{\boldsymbol{\beta}}$ solves optimization problem (8) for $\lambda = \lambda_2$ by verifying that it satisfies the corresponding KKT conditions. Note that if $\|\mathbf{h}_K\|_\infty < \min_K |\beta_j|$, then $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ have the same sign pattern. If this is the case, then the KKT equality conditions become $X_K^T(-X_K\mathbf{h}_K + \boldsymbol{\epsilon}) = \lambda_2\mathbf{s}$ and are satisfied by the definition of \mathbf{h} . The KKT inequality conditions become $\|X_{K^c}(-X_K\mathbf{h}_K + \boldsymbol{\epsilon})\|_\infty < \lambda_2$. All that is left to do is write a set of requirements on m_1, M_2 and λ_1 under which the last inequality holds together with $\|\mathbf{h}_K\|_\infty < \min_K |\beta_j|$.

Note that the term involving $\boldsymbol{\epsilon}$ in the expression for \mathbf{h}_K is V , defined in Lemma 1, and the corresponding term in the KKT inequality conditions is exactly $\lambda_2 W$. Consequently, the desired inequalities follow from

$$\|\Sigma_{K^c K} \Sigma_{KK}^{-1} \mathbf{s}\|_\infty + \lambda_2 W < 1 \quad \text{and} \quad \lambda_2 \|\Sigma_{KK}^{-1} \mathbf{s}\|_\infty + V < \min_{j \in K} |\beta_j|.$$

Taking into account conditions (2) and (7), the requirements simplify to

$$\|\Sigma_{K^c K} \Sigma_{KK}^{-1} \mathbf{s}\|_\infty < 1 - \frac{\xi}{2} \quad \text{and} \quad \|\Sigma_{KK}^{-1} \mathbf{s}\|_\infty < (c_1/c_2)\sqrt{S} - 4\sqrt{2}\sigma. \quad (9)$$

Let $\|\cdot\|$ denote the operator norm of a matrix and recall that $\|\Sigma_{KK}^{-1}\| \leq 2$, as we mentioned after Lemma 1. Hence the second inequality in the above display is satisfied for $c_1 = c_2(2 + 4\sqrt{2}\sigma/\sqrt{S})$. It is only left to establish the first inequality. Note that $\Sigma_{KK}^{-1} \mathbf{s} = \Sigma_{KK}^{-1} E \mathbf{s}$, where $E = \text{diag}(|\mathbf{s}|)$ and the $|\cdot|$ operation is understood componentwise. It follows that

$$\|\Sigma_{K^c K} \Sigma_{KK}^{-1} \mathbf{s}\|_\infty \leq \mu S q_2^{1/2} \|\Sigma_{KK}^{-1} E\|.$$

Define $A = I - \Sigma_{KK}$ and note that if inequality $\|A\| < 1$ holds, then $\Sigma_{KK}^{-1} = \sum_{j=0}^{\infty} A^j$ and $\|\Sigma_{KK}^{-1} E\| \leq 1 + \|AE\|/(1 - \|A\|)$. Arguing as Zhao and Yu (2006) in the proof of their Corollary 2, but without explicitly specifying the correlation bound, we get $\|A\| \leq \mu(S - 1)$ and $\|AE\| \leq \mu(q_2 S - 1)$. Consequently, inequality $\mu < 1/[S - 1]$ implies that the needed bound $\|A\| < 1$ holds, and the preceding argument goes

through. Conclude that

$$\|\Sigma_{K^c K} \Sigma_{K^c K}^{-1} \mathbf{s}\|_\infty \leq \mu S q_2^{1/2} [1 - \mu q_1 S] / [1 - \mu(S - 1)]. \quad (10)$$

The right hand side is an increasing function of μ for $q_1 \leq (S - 1)/S$, which covers all the nontrivial cases. A direct calculation shows that the right-hand side is below $1 - \xi$ for $\mu = (1 - \xi) / [(2 - q_1)S]$. This establishes the first inequality in (3).

Note that we could improve the bound $\mu < \mu_{FL}(1 - \xi)$ for very small values of q_1 , as the right-hand side of (10) is strictly below $1 - \xi$ for each $\mu < (1 - \xi) / [(1 + q_2^{1/2})S - 1]$.

4 Proof of Lemma 2.2

For simplicity of the notation we replace expressions K_1, K_2 , and K_1^c in various subscripts by 1, 2, and 1^c , respectively.

Consider a $\hat{\boldsymbol{\beta}}$ that satisfies $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \mathbf{h}_1$ with $\mathbf{h}_1 = \Sigma_{11}^{-1}(-\lambda \text{sign}(\boldsymbol{\beta}_1) + \Sigma_{12} \boldsymbol{\beta}_2 + X_1^T \boldsymbol{\epsilon})$ and $\hat{\boldsymbol{\beta}}_{1^c} = 0$. It is sufficient to show that $\hat{\boldsymbol{\beta}}$ solves the Lasso problem for the tuning parameter $\lambda = \lambda_1$, which we will do by verifying the corresponding KKT conditions. Note that if $\|\mathbf{h}_1\|_\infty < m_1$, then $\text{sign}(\hat{\boldsymbol{\beta}}_1) = \text{sign}(\boldsymbol{\beta}_1)$. Thus the KKT equality conditions become $X_1^T(-X_1 \mathbf{h}_1 + X_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}) = \lambda_1 \text{sign}(\boldsymbol{\beta}_1)$ and are satisfied by the definition of \mathbf{h}_1 . The KKT inequality conditions become $\|X_{1^c}(-X_1 \mathbf{h}_1 + X_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon})\|_\infty < \lambda_1$. All that is left to do is write a set of requirements on m_1, M_2 and λ_1 under which the last inequality holds together with $\|\mathbf{h}_1\|_\infty < m_1$.

Note that the term involving $\boldsymbol{\epsilon}$ in the KKT inequality conditions is exactly $\lambda_2 \tilde{W}$, defined in Lemma 1, and the corresponding term in the expression for \mathbf{h}_1 is \tilde{V} . The desired inequalities follow from

$$\begin{aligned} \|\Sigma_{1^c 1} \Sigma_{11}^{-1} \lambda_1 \mathbf{s}_1\|_\infty + \|(\Sigma_{1^c 2} - \Sigma_{1^c 1} \Sigma_{11}^{-1} \Sigma_{12}) \boldsymbol{\beta}_2\|_\infty + \lambda_2 \tilde{W} &< \lambda_1 \quad \text{and} \\ \|\Sigma_{11}^{-1} \lambda_1 \mathbf{s}_1\|_\infty + \|\Sigma_{11}^{-1} \Sigma_{12} \boldsymbol{\beta}_2\|_\infty + \tilde{V} &< m_1. \end{aligned}$$

Using $\lambda_1 > \lambda_2$ and linear algebra arguments along the lines of those in the previous section we can conclude that it suffices to check

$$\begin{aligned} \lambda_1(1 - \frac{\xi}{2}) + [M_2 S \mu(1 + [1 - 2q_1]S\mu + \mu)] / [1 - q_1 S \mu + \mu] &\leq \lambda_1 \quad \text{and} \\ (q_1 S)^{1/2} [\lambda_1 + M_2 q_2 S \mu] / [1 - q_1 S \mu + \mu] + \lambda_2 4\sqrt{2}\sigma &< m_1. \end{aligned}$$

Rearranging the terms yields that the above conditions are satisfied if $m_1/M_2 > c_3 S^{1/2}$ and $\lambda_1 = c_4 M_2$, where

$$\begin{aligned} c_3 &\geq q_1^{1/2} [c_4 + q_2 S \mu] / [1 - q_1 S \mu + \mu] + 4\sqrt{2} \sigma c_2 / (c_1 \sqrt{S}) \quad \text{and} \\ c_4 &\geq (2/\xi) S \mu (1 + [1 - 2q_1] S \mu + \mu) / [1 - q_1 S \mu + \mu]. \end{aligned}$$

The right hand sides in the above display can be easily bounded by constants using the definition of μ_{FL} in display (3). We complete the proof by taking c_3 and c_4 as these constant bounds.

References

- Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recover using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.
- Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541–2563.