Supplementary materials for *Statistical Estimation and Testing via the Sorted ℓ_1 Norm*

Małgorzata Bogdan* Ewout van den Berg† Weijie Su‡ Emmanuel J. Candès§

October 2013

Abstract

In this note we give a proof showing that even though the number of false discoveries and the total number of discoveries are not continuous functions of the parameters, the formulas we obtain for the false discovery proportion (FDP) and the power, namely, (B.3) and (B.4) in the paper *Statistical Estimation and Testing via the Sorted ℓ_1 Norm* are mathematically valid. We recall that these formulas are derived from [1, Theorem 1.5].

Consider the linear model

$$y = X\beta + z,$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta \in \mathbb{R}^p$ is the unknown parameter of interest and $z \in \mathbb{R}^n$ is a vector of i.i.d. standard Gaussian independent of X and β. The lasso estimate $\hat{\beta}$ with penalty $\lambda > 0$ is the solution to

$$\min_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \lambda\|b\|_{\ell_1}. \tag{1}$$

As in the main paper, we let $\varphi_V(x, y) = 1(x \neq 0)1(y = 0)$ and $\varphi_R(x, y) = 1(x \neq 0)$ so that the number V of false discoveries is equal to

$$V = \sum_i \varphi_V(\hat{\beta}_i, \beta_i),$$

and the number R of discoveries is equal to

$$R = \sum_i \varphi_R(\hat{\beta}_i, \beta_i).$$

We can now state the main result in this note, which generalizes [1, Theorem 1.5].

Theorem 1. *Suppose that X is an $n \times p$ Gaussian design matrix with i.i.d. $\mathcal{N}(0, \frac{1}{n})$ entries, β_i's are non-degenerate i.i.d. random variables with bounded second moment independent of X. Below, Θ is a variable with the same distribution as β_i. As $p \to \infty$ and $n/p \to \delta > 0$, the lasso solution $\hat{\beta}$ obeys*

$$\text{FDP} \equiv \frac{\sum_{i=1}^p \varphi_V(\hat{\beta}_i, \beta_i)}{\max\{\sum_{i=1}^p \varphi_R(\hat{\beta}_i, \beta_i), 1\}} \xrightarrow{P} \frac{2\mathbb{P}(\Theta = 0)\Phi(-\alpha)}{\mathbb{P}(|\Theta + \tau Z| > \alpha\tau)},$$

where $\tau > 0$ and $\alpha > \alpha_{\min}$ are the unique solutions to

$$\tau^2 = 1 + \frac{1}{\delta}\mathbb{E}\Big(\eta_{\alpha\tau}(\Theta + \tau Z) - \Theta\Big)^2,$$
$$\lambda = \Big(1 - \frac{1}{\delta}\mathbb{P}(|\Theta + \tau Z| > \alpha\tau)\Big)\alpha\tau.$$

Here η is the soft-thresholding operator defined as $\eta_t(x) = \text{sign}(x)(|x| - t)_+$ and Z is a standard Gaussian independent of Θ.

*Departments of Mathematics and Computer Science, Wrocław University of Technology and Jan Długosz University, Poland
†IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.
‡Department of Statistics, Stanford University, Stanford, CA 94305
§Departments of Statistics and of Mathematics, Stanford University, Stanford, CA 94305

Before presenting the proof, we give three lemmas.

Lemma 1. *Suppose*

$$\Sigma = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{pmatrix}$$

is a positive definite matrix with all eigenvalues larger than or equal to 1, where $\Sigma_{1,1}$ is a scalar. Then we have

$$\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} \geq 1.$$

Proof. The condition $\Sigma \succeq I$, where I is the identity matrix with the same size as Σ, is equivalent to $0 \prec \Sigma^{-1} \preceq I$. So by the Schur complement property, the $(1,1)$ entry of Σ^{-1} satisfies

$$0 < \Sigma^{-1}(1,1) = (\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1})^{-1} \leq 1,$$

which gives $\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1} \geq 1$ as desired. □

Lemma 2. *Suppose Y is a p-dimensional vector distributed as $\mathcal{N}(\mu, \Sigma)$, where $\Sigma \succeq \sigma^2 I$. For any $\epsilon \in (0,1)$, there exists a constant $c = c(\epsilon) > 0$ such that for any $h > 0$,*

$$\mathbb{P}(\textit{at least } \epsilon p \textit{ components of } Y \textit{ are in } (-h,h)) \leq \min(1, ch^{\epsilon}\sigma^{-\epsilon})^p.$$

Proof. By Lemma 1, the variance of $Y_i|Y_{-i}$, where the subscript $_{-i}$ denotes all the components except the ith, is larger than or equal to σ^2. So we have $\mathbb{P}(Y_i \in (-h,h)|Y_{-i}) \leq \min(1, \frac{2h}{\sqrt{2\pi}\sigma})$ almost surely since the normal density function ϕ is bounded by $\frac{1}{\sqrt{2\pi}}$. Denote by ξ_i i.i.d. Bernoulli variables independent of Y with $\mathbb{P}(\xi_i = 1) = \tilde{p} \equiv \min(1, \frac{2h}{\sqrt{2\pi}\sigma})$. And denote by $\zeta_i = 1(Y_i \in (-h,h))$. Since

$$\mathbb{P}(\zeta_1 = 1|\zeta_2, \ldots, \zeta_p) \leq \mathbb{P}(\xi_1 = 1|\zeta_2, \ldots, \zeta_p), \text{ a.s.},$$

we have

$$\mathbb{P}(\zeta_1 + \zeta_2 + \ldots + \zeta_p \geq \epsilon p) \leq \mathbb{P}(\xi_1 + \zeta_2 + \ldots + \zeta_p \geq \epsilon p).$$

Using similar arguments we reach the conclusion

$$\mathbb{P}(\xi_1 + \xi_2 + \ldots + \xi_k + \zeta_{k+1} + \ldots + \zeta_p \geq \epsilon p) \leq \mathbb{P}(\xi_1 + \xi_2 + \ldots + \xi_k + \xi_{k+1} + \zeta_{k+2} + \ldots + \zeta_p \geq \epsilon p),$$

which holds for $k = 1, 2, \ldots, p-1$. Therefore,

$$\begin{aligned} \mathbb{P}(\text{at least } \epsilon p \text{ components of } Y \text{ are in } (-h,h)) &= \mathbb{P}(\zeta_1 + \zeta_2 + \ldots + \zeta_p \geq \epsilon p) \\ &\leq \mathbb{P}(\xi_1 + \zeta_2 + \ldots + \zeta_p \geq \epsilon p) \\ &\leq \mathbb{P}(\xi_1 + \xi_2 + \zeta_3 + \ldots + \zeta_p \geq \epsilon p) \\ &\ldots \\ &\leq \mathbb{P}(\xi_1 + \xi_2 + \ldots + \xi_p \geq \epsilon p). \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}(\sum_{i=1}^{p} \xi_i \geq \epsilon p) &\leq \sum_{i \geq \epsilon p} \binom{n}{i} \tilde{p}^i \\ &\leq \tilde{p}^{\epsilon p} \sum_{i \geq \epsilon p} \binom{n}{i} \\ &\leq \tilde{p}^{\epsilon p} 2^p \\ &\leq (\frac{2^{1+\epsilon/2}}{\pi^{\epsilon/2}})^p h^{\epsilon p} \sigma^{-\epsilon p}. \end{aligned}$$

□

Lemma 3. *In the same setting as Theorem 1, denote by $\mathcal{A} \subset \{1, \dots, p\}$ the active set of the lasso solution $\hat{\beta}$ of* (1)*. Then there exists a constant $\rho > 0$ such that*

$$\lim_{p\to\infty} \mathbb{P}(|\mathcal{A}| < \rho p) = 0.$$

Proof. Define $\varphi(x, y) = \min(|x|, 1)$, which is pseudo-Lipschitz. It follows from [1, Theorem 1.5] that

$$\lim_{p\to\infty} \frac{\sum_{i=1}^p \min(|\hat{\beta}_i|, 1)}{p} = \mathbb{E}\min(|\eta_{\alpha\tau}(\Theta + \tau Z)|, 1) > 0$$

in probability. Note that

$$\frac{|\mathcal{A}|}{p} \geq \frac{\sum_{i=1}^p \min(|\hat{\beta}_i|, 1)}{p}.$$

So for any $\rho < \mathbb{E}\min(|\eta_{\alpha\tau}(\Theta + \tau Z)|, 1)$ we have $\lim_{p\to\infty} \mathbb{P}(|\mathcal{A}| < \rho p) = 0$. □

Proof of Theorem 1. To go around the discontinuity of φ_V, we define a series of pseudo-Lipschitz continuous functions $\varphi_{V,h}(x, y) = (1 - Q(x/h))Q(y/h)$, where $Q(x) = \max(1 - |x|, 0)$ for $h > 0$. Therefore, by [1, Theorem 1.5],

$$\lim_{p\to\infty} \frac{1}{p} \sum_{i=1}^p \varphi_{V,h}(\hat{\beta}_i, \beta_i) = \mathbb{E}\varphi_{V,h}(\eta_{\alpha\tau}(\Theta + \tau Z), \Theta) \tag{2}$$

in probability. Since

$$|\varphi_{V,h}(x, y) - \varphi_V(x, y)| \leq 1(0 < |x| < h) + 1(0 < |y| < h), \tag{3}$$

so that for any $\epsilon > 0$,

$$\mathbb{P}\Big(|\frac{1}{p}\sum_{i=1}^p \varphi_V(\hat{\beta}_i, \beta_i) - \frac{1}{p}\sum_{i=1}^p \varphi_{V,h}(\hat{\beta}_i, \beta_i)| > \epsilon\Big) \leq \mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^p 1(0 < |\hat{\beta}_i| < h) > \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^p 1(0 < |\beta_i| < h) > \frac{\epsilon}{2}\Big).$$

By the weak Law of Large Numbers we have

$$\lim_{h\to 0}\lim_{p\to\infty} \mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^p 1(0 < |\beta_i| < h) > \frac{\epsilon}{2}\Big) = 0.$$

So if we additionally have

$$\lim_{h\to 0}\limsup_{p\to\infty} \mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^p 1(0 < |\hat{\beta}_i| < h) > \frac{\epsilon}{2}\Big) = 0 \tag{4}$$

for any $\epsilon > 0$, we would obtain

$$\begin{aligned}
\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^p \varphi_V(\hat{\beta}_i, \beta_i) &\overset{P}{=} \lim_{h\to 0}\lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^p \varphi_{V,h}(\hat{\beta}_i, \beta_i) \\
&\overset{P}{=} \lim_{h\to 0} \mathbb{E}\varphi_{V,h}(\eta_{\alpha\tau}(\Theta + \tau Z), \Theta) \\
&= \mathbb{E}\varphi_V(\eta_{\alpha\tau}(\Theta + \tau Z), \Theta),
\end{aligned} \tag{5}$$

where the last equality comes from applying the dominated convergence theorem to $\varphi_{V,h} \to \varphi_V$.

We prove (4) to complete the proof of the theorem. Denote the active set of the Lasso solution by $\mathcal{A} \subset \{1, \dots, p\}$. Then partial KKT conditions give

$$X_{\mathcal{A}}^T(y - X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) = \tilde{\lambda}_{\mathcal{A}},$$

where $\tilde{\lambda}_{\mathcal{A}}$ is a vector with each component being λ or $-\lambda$ depending the signs of $\hat{\beta}_{\mathcal{A}}$. Note that $X_{\mathcal{A}}^T X_{\mathcal{A}}$ is invertible with probability one because $|\mathcal{A}| \leq n$. So we may write the solution as

$$\begin{aligned}
\hat{\beta}_{\mathcal{A}} &= (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}(X_{\mathcal{A}}^T y - \tilde{\lambda}_{\mathcal{A}}) \\
&= (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1}(X_{\mathcal{A}}^T X\beta - \tilde{\lambda}_{\mathcal{A}}) + (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T z.
\end{aligned}$$

Now for any subset $\mathcal{D} \subset \{1, \ldots, p\}$ with $|\mathcal{D}| \le n$, and $\tilde{\lambda}$ of length $|\mathcal{D}|$ with each component being $\pm\lambda$, define

$$\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}} = (X^T_{\mathcal{D}} X_{\mathcal{D}})^{-1}(X^T_{\mathcal{D}} X\beta - \tilde{\lambda}) + (X^T_{\mathcal{D}} X_{\mathcal{D}})^{-1} X^T_{\mathcal{D}} z.$$

If $\mathcal{D} = \mathcal{A}$ and $\tilde{\lambda} = \tilde{\lambda}_{\mathcal{A}}$, $\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}}$ coincides with $\hat{\beta}_{\mathcal{A}}$.

By Lemma 3, there is a constant $\rho > 0$ such that $\mathbb{P}(|\mathcal{A}| < \rho p) = o(1)$. For any $\epsilon' > 0$, by the union bound we have

$$\mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^{p} 1(0 < |\hat{\beta}_i| < h) > \epsilon\Big) \le \sum_{\rho p \le |\mathcal{D}| \le \min(p,n)} \sum_{|\tilde{\lambda}| = |\mathcal{D}|} \mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^{|\mathcal{D}|} 1(|\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}}(i)| < h) > \epsilon, \sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon'\Big) + \mathbb{P}\Big(\sigma_{\max}(X) \ge \delta^{-1/2} + 1 + \epsilon'\Big) + \mathbb{P}(|\mathcal{A}| < \rho p). \quad (6)$$

By [2], we have

$$\mathbb{P}(\sigma_{\max}(X) \ge \delta^{-1/2} + 1 + \epsilon') = o(1). \quad (7)$$

Now we estimate

$$\mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^{|\mathcal{D}|} 1(|\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}}(i)| < h) > \epsilon, \sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon'\Big).$$

Conditionally on X, $\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}}$ is distributed as Gaussian with mean $(X^T_{\mathcal{D}} X_{\mathcal{D}})^{-1}(X^T_{\mathcal{D}} X\beta - \tilde{\lambda})$ and covariance $(X^T_{\mathcal{D}} X_{\mathcal{D}})^{-1}$. On the event $\sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon'$, all the eigenvalues of $(X^T_{\mathcal{D}} X_{\mathcal{D}})^{-1}$ are larger than $\sigma'^2 \triangleq (\delta^{-1/2} + 1 + \epsilon')^{-2}$. So by Lemma 2,

$$\begin{aligned}
&\mathbb{P}\Big(\frac{1}{p}\sum_{i=1}^{|\mathcal{D}|} 1(|\hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}}(i)| < h) > \epsilon, \sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon' | X\Big) \\
&\quad \le \mathbb{P}(\text{at least } \epsilon|\mathcal{D}| \text{ components of } \hat{\beta}^{\tilde{\lambda}}_{\mathcal{D}} \text{ are in } (-h, h), \sigma_{\max}(X) < \delta^{-1/2} + 1 + \epsilon' | X) \\
&\quad \le \min(1, ch^{\epsilon}\sigma'^{-\epsilon})^{|\mathcal{D}|} \\
&\quad \le \min(1, ch^{\epsilon}\sigma^{-\epsilon})^{\rho p} \\
&\quad \le (ch^{\epsilon}\sigma'^{-\epsilon})^{\rho p}.
\end{aligned}$$

Together with (6) and (7) this gives

$$\begin{aligned}
&\mathbb{P}(\frac{1}{p}\sum_{i=1}^{p} 1(0 < |\hat{\beta}_i| < h) > \epsilon) \\
&\le \sum_{\rho p \le |\mathcal{D}| \le \min(p,n)} \sum_{|\tilde{\lambda}| = |\mathcal{D}|} (ch^{\epsilon}\sigma'^{-\epsilon})^{\rho p} + o(1) + o(1) \\
&\le 2^p 2^p (ch^{\epsilon}\sigma'^{-\epsilon})^{\rho p} + o(1) \\
&= (4c^{\rho} h^{\rho\epsilon}\sigma'^{-\rho\epsilon})^p + o(1),
\end{aligned} \quad (8)$$

which proves (4) by choosing h sufficiently small such that $4c^{\rho} h^{\rho\epsilon}\sigma'^{-\rho\epsilon} < 1$. Consequently (5) is also proved.

This gives

$$\lim_{p\to\infty} \frac{V}{p} \overset{P}{=} \mathbb{E}\varphi_V(\eta_{\alpha\tau}(\Theta + \tau Z), \Theta). \quad (9)$$

Similarly, defining $\varphi_{R,h}(x, y) = 1 - Q(x/h)$, we can also establish

$$\lim_{p\to\infty} \frac{R}{p} \overset{P}{=} \mathbb{E}\varphi_R(\eta_{\alpha\tau}(\Theta + \tau Z), \Theta). \quad (10)$$

Combining (9) and (10) gives

$$\begin{aligned}\text{FDP} = \frac{V}{\max(R,1)} &\xrightarrow{P} \frac{\mathbb{E}\varphi_V(\eta_{\alpha\tau}(\Theta+\tau Z),\Theta)}{\mathbb{E}\varphi_R(\eta_{\alpha\tau}(\Theta+\tau Z),\Theta)} \\ &= \frac{2\mathbb{P}(\Theta=0)\Phi(-\alpha)}{\mathbb{P}(|\Theta+\tau Z|>\alpha\tau)}.\end{aligned}$$

□

As a byproduct of the proof of Theorem 1, we have

Corollary 1. *With the same setting as Theorem 1, the empirical power of the lasso solution $\hat{\beta}$ of* (1) *obeys*

$$\text{Power} \equiv \frac{\#\{1\le i\le p : \hat{\beta}_i\neq 0, \beta_i\neq 0\}}{\|\beta\|_{\ell_0}} \xrightarrow{P} \mathbb{P}(|\Theta+\tau Z|>\alpha\tau|\Theta\neq 0).$$

Proof of Corollary. We have

$$\begin{aligned}\text{Power} &= \frac{R-V}{\#\{1\le i\le p:\beta_i\neq 0\}} \\ &= \frac{(R-V)/p}{\#\{1\le i\le p:\beta_i\neq 0\}/p} \\ &\xrightarrow{P} \frac{\mathbb{E}\varphi_R(\eta_{\alpha\tau}(\Theta+\tau Z),\Theta)-\mathbb{E}\varphi_V(\eta_{\alpha\tau}(\Theta+\tau Z),\Theta)}{\mathbb{P}(\Theta\neq 0)} \\ &= \frac{\mathbb{P}(\Theta\neq 0,|\Theta+\tau Z|>\alpha\tau)}{\mathbb{P}(\Theta\neq 0)} \\ &= \mathbb{P}(|\Theta+\tau Z|>\alpha\tau|\Theta\neq 0).\end{aligned}$$

□

References

[1] M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[2] S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980.