Supplementary Material for "Regularization for Unsupervised Deep Neural Nets"

Baiyang Wang, Diego Klabjan

Supplement for Section 5.2

Unlike a deep belief network (DBN) which trains a directed model, a deep Boltzmann machine (DBM) trains an undirected graphical model. The major difference is that in the training procedure for a DBM, both the visible and hidden nodes in the deepest hidden layer are doubled. The weights for the doubled nodes are tied to the original weights. Also, $E(h^L|v)$ may be added to the original features.

For a replicated softmax model (RSM), the energy function is

$$\begin{cases} E(v,h) = -b^T v - C \cdot c^T h - h^T W v, \\ 1^T v = C,\ v \in \mathbb{N}^J,\ h \in \{0,1\}^I. \end{cases}$$

For a Gaussian RBM, the energy function is

$$\begin{cases} E(v,h) = \sum_{j=1}^J \frac{(v_j - a_j)^2}{2\sigma_j^2} - c^T h - h^T W(v/\sigma), \\ v \in \mathbb{R}^J,\ h \in \{0,1\}^I. \end{cases}$$

where v/σ is element-wise.

RSMs and Gaussian RBMs are simple extensions of an RBM to count outcomes and real-valued outcomes.

Supplement for Sections 4 - 5.1

Theorem 1. (adapted from Theorem 5.7 in [6])

Let θ be a vector. Let $M^{(N)}(\vartheta)$ be a random variable parametrized by ϑ and let M be a fixed function of ϑ such that for every $\epsilon > 0$,

$$\sup_{\vartheta \in \Theta} |M^{(N)}(\vartheta) - M(\vartheta)| \xrightarrow{P} 0, \quad \sup_{\vartheta : \|\vartheta - \theta\| \geq \epsilon} M(\vartheta) < M(\theta). \tag{1}$$

Then for any sequence of random variables $\{\hat{\theta}^{(N)}\}_{N=1}^{\infty}$ such that $M^{(N)}(\hat{\theta}^{(N)}) \geq M^{(N)}(\theta) - o_P(1)$ a.s., we have $\hat{\theta}^{(N)} \xrightarrow{P} \theta$. □

Lemma 1. (adapted from page 46 in [6])

For $M^{(N)}(\vartheta) = \frac{1}{N} \sum_{n=1}^N m(v^{(n)}|\vartheta)$, where $v^{(n)}$ is defined in Section 4, and $m(v^{(n)}|\vartheta)$ is a function of both $v^{(n)}$ and ϑ, if $\vartheta \in \Theta$ which is compact, the functions $\{m(v|\vartheta) : \vartheta \in \Theta\}$ are continuous for every v, and they are dominated by an integrable function, then for $M(\vartheta)$ such that $M^{(N)}(\vartheta) \xrightarrow{P} M(\vartheta)$ for each $\vartheta \in \Theta$, we have $\sup_{\vartheta \in \Theta} |M^{(N)}(\vartheta) - M(\vartheta)| \xrightarrow{P} 0$. □

Theorem 2. (adapted from page 1361 in [4]; see [2, 7] also)

Let u be another parametrization of ϑ such that $\vartheta(u) = \theta + u/\sqrt{N}$, and $u(\vartheta) = \sqrt{N}(\vartheta - \theta)$. If $V^{(N)}(u) \xrightarrow{d} V(u) = -2u^T W + u^T C u + \infty \cdot \sum_{d > D_0}^D 1_{u_d \neq 0}$ for each $u \in \mathbb{R}^D$ such that $\vartheta(u) \in \Theta$, where $V^{(N)}(u)$ is a random variable parametrized by u, W is a random variable and C is a constant matrix, then for $\hat{\theta}^{(N)} = \arg\min_\vartheta V^{(N)}(u(\vartheta))$, $\sqrt{N}(\hat{\theta}^{(N)} - \theta) \xrightarrow{d} \arg\min_u V(u)$. □

Proof of Proposition 1.

(a) By assumption we have

$$\begin{aligned} M^{(N)}(\vartheta) &= -\frac{1}{N}\tilde{l}^{(N)}(\vartheta) \\ &= -\frac{1}{N}l^{(N)}(\vartheta) + \sum_{d=1}^{D}\frac{\mu_d^{(N)}}{N}|\vartheta_d| + \sum_{d=1}^{D}\frac{\lambda_d^{(N)}}{N}\vartheta_d^2 \\ &\xrightarrow{P} -E[l(v|\theta)] = M(\vartheta). \end{aligned} \tag{2}$$

Here, $l^{(N)}(\vartheta)$ denotes the log-likelihood with parameter $\vartheta \in \Theta$ and data examples $v^{(1)}, \ldots, v^{(N)}$, and $l(v|\theta) = \log P(v|\theta)$. Verifying the conditions for Theorem 1, indeed, $-\frac{1}{N}\tilde{l}^{(N)}(\vartheta)$ is a random variable parametrized by ϑ and $-E[l(v|\vartheta)]$ is a fixed function for $\vartheta \in \Theta$.

For the first condition in (1), the regularization terms in (2) converge to zero uniformly because Θ is compact, so we only need the uniform convergence of $-\frac{1}{N}l^{(N)}(\vartheta)$. This is established by Lemma 1. Indeed $m(v|\vartheta) = -l(v|\vartheta)$ is continuous for every v. Because Θ is compact and $v \in \{0,1\}^J$, $-l(v|\vartheta)$ is bounded and thus dominated by an integrable function. Therefore, Lemma 1 holds due to (2) and the first condition in (1) is satisfied.

For the second condition in (1), we note that $-E[l(\theta)] \leq -E[l(\vartheta)]$ from Jensen's equality, and the equality is reached only if $P(v|\theta) = P(v|\vartheta)$. Because we have assumed that the model is identifiable, this implies that $\vartheta = \theta$. Therefore, the second condition in (1) is satisfied.

Because we assume $\tilde{\theta}^{(N)}$ minimizes $M^{(N)}(\vartheta)$, applying Theorem 1, $\tilde{\theta}^{(N)} \xrightarrow{P} \theta$.

(b) We generally follow [7]. Let θ be defined as in the paper, $u \in \mathbb{R}^D$ be a fixed constant, and let $\vartheta(u) = \theta + u/\sqrt{N}$. We have the following Lagrange form of Taylor's expansion for some $\dot{\vartheta}$ lying between $\vartheta(u)$ and θ,

$$\begin{aligned} V^{(N)}(u) &= \frac{1}{N}[\tilde{l}^{(N)}(\vartheta(u)) - l^{(N)}(\theta)] \\ &= -\frac{1}{\sqrt{N}}\frac{\partial}{\partial\vartheta}l^{(N)}(\theta)u - \frac{u^T}{2N}\frac{\partial^2}{\partial\vartheta^2}l^{(N)}(\dot{\vartheta})u + \sum_{d=1}^{D}\mu_d^{(N)}\left(\left|\theta_d + \frac{u_d}{\sqrt{N}}\right| - |\theta_d|\right) \\ &\quad + \sum_{d=1}^{D}\lambda_d^{(N)}\frac{u_d}{\sqrt{N}}\left(2\theta_d + \frac{u_d}{\sqrt{N}}\right) \\ &\xrightarrow{d} N(0, I(\theta))u + \frac{1}{2}u^T I(\theta)u + \infty \cdot \sum_{d=1}^{D} 1_{u_d \neq 0,\ \theta_d = 0} = V(u). \end{aligned} \tag{3}$$

Equation (3) holds because $\frac{1}{\sqrt{N}}\frac{\partial}{\partial\vartheta}l^{(N)}(\theta) \xrightarrow{d} N(0, I(\theta))$, $\frac{1}{N}\frac{\partial^2}{\partial\vartheta^2}l^{(N)}(\theta) \xrightarrow{P} I(\theta)$, and the two regularization terms converge accordingly. The only remaining result needed for (3) is then $\frac{1}{N}[\frac{\partial^2}{\partial\vartheta^2}l^{(N)}(\dot{\vartheta}) - \frac{\partial^2}{\partial\vartheta^2}l^{(N)}(\theta)] \xrightarrow{P} 0$. This can be shown by taking another Taylor's expansion, and from the fact that $\frac{\partial^3}{\partial\vartheta^3}l(v|\vartheta)$ is bounded with respect to v and ϑ. Therefore, asymptotic normality follows from Theorem 2 with equations (3). Moreover, we also have $\sqrt{N}\tilde{\theta}_d \xrightarrow{P} 0$ if $\theta_d = 0$.

To prove the rest, we only need to show $P(\tilde{\theta}_d^{(N)} \neq 0) \to 0$ where $\theta_d = 0$. We note that if $\tilde{\theta}_d^{(N)} \neq 0$, then $\frac{\partial M^{(N)}}{\partial\vartheta_d}(\tilde{\theta}^{(N)})$ exists and should be 0, otherwise we can get a lower $M^{(N)}(\vartheta)$. Equivalently,

$$-\frac{1}{\sqrt{N}}\frac{\partial l^{(N)}}{\partial\vartheta_d}(\tilde{\theta}^{(N)}) + \frac{2\lambda_d^{(N)}}{\sqrt{N}}\tilde{\theta}_d^{(N)} \pm \frac{\mu_d^{(N)}}{\sqrt{N}} = 0. \tag{4}$$

By assumption, $2\lambda_d^{(N)}\tilde{\theta}_d^{(N)}/\sqrt{N} \to 0$, $\mu_d^{(N)}/\sqrt{N} \to \infty$. We have

$$\frac{1}{\sqrt{N}}\frac{\partial l^{(N)}}{\partial \vartheta_d}(\tilde{\theta}^{(N)}) = \frac{1}{\sqrt{N}}\frac{\partial l^{(N)}}{\partial \vartheta_d}(\theta) + \sum_{d'=1}^{D}\sqrt{N}(\tilde{\theta}_{d'}^{(N)} - \theta_{d'}) \cdot \frac{1}{N}\frac{\partial^2 l^{(N)}}{\partial \vartheta_d \vartheta_{d'}}(\ddot{\vartheta}) = O_P(1). \tag{5}$$

Equation (5) holds because $\frac{1}{\sqrt{N}}\frac{\partial l^{(N)}}{\partial \vartheta_d}(\theta)$ and $\sqrt{N}(\tilde{\theta}_{d'}^{(N)} - \theta_{d'})$ converge to normal random variables, and $\frac{1}{N}\frac{\partial^2 l^{(N)}}{\partial \vartheta_d \vartheta_{d'}}(\ddot{\vartheta})$ is bounded by $\sup_{\vartheta\in\Theta, v\in\{0,1\}^J}\left|\frac{\partial^2 l(v|\vartheta)}{\partial\vartheta_d\partial\vartheta_{d'}}\right|$. Therefore equation (4) hold with probability tending to zero, and we have $P(\hat{A}^{(N)} = A) \to 1$. □

Proof of Proposition 2.

Let $m^{(N)}$ represent the mask over each component of ϑ for N data examples. From any $p^{(N)}$, $m^{(N)}$ is a random variate such that $m^{(N)} \sim Bernoulli(p^{(N)})$ element-wise. Then because $p^{(N)} \to 1$ (1 is multi-dimensional where appropriate), for any $m \neq 1$, $P(m^{(N)} = m) \to 0$ as $N \to \infty$. This is true for both original and partial Dropout/DropConnect. Then

$$\begin{aligned}
-\frac{1}{N}\tilde{l}^{(N)}(\vartheta) &= -\frac{1}{N}\sum_{n=1}^{N} E_{m^{(N)}} l(v^{(n)}|\vartheta * m^{(N)}) \\
&= -\frac{1}{N}\sum_{n=1}^{N}\left\{l(v^{(n)}|\vartheta) + \left[E_{m^{(N)}} l(v^{(n)}|\vartheta * m^{(N)}) - l(v^{(n)}|\vartheta)\right]\right\} \\
&= -\frac{1}{N}l^{(N)}(\vartheta) - \sum_{m\neq 1} P(m^{(N)} = m) \cdot \frac{1}{N}\sum_{n=1}^{N}\left[l(v^{(n)}|\vartheta * m^{(N)}) - l(v^{(n)}|\vartheta)\right] \\
&= -E[l(\vartheta)] + o_P(1) + \sum_{m\neq 1} o(1)\cdot(C + o_P(1)) \xrightarrow{P} -E[l(\vartheta)].
\end{aligned} \tag{6}$$

For uniform convergence, we only need to show that $\sup_{\vartheta\in\Theta}\frac{1}{N}\sum_{n=1}^{N}[l(v^{(n)}|\vartheta * m) - l(v^{(n)}|\vartheta)] = O_P(1)$, which is equivalent to $\sup_{\vartheta\in\Theta}\frac{1}{N}\sum_{n=1}^{N} l(v^{(n)}|\vartheta * m) = O_P(1)$. To do this, we apply the conditions for Lemma 1 again, noting that

$$\sup_{\vartheta\in\Theta}\left|\frac{1}{N}\sum_{n=1}^{N} l(v^{(n)}|\vartheta * m) - E[l(v|\vartheta * m)]\right| \to 0 \tag{7}$$

and $E[l(v|\vartheta * m)]$ is bounded for ϑ because of continuity and compactness of Θ. Therefore, uniform convergence holds, and furthermore, the conclusion holds, same as in Proposition 1(a). □

Theorem 3. (adapted from Theorem 1 in [5])

Let $\hat{\theta}^{(N)}$ be as in Section 4.

If A1) there exists a compactification $\bar{\Theta}$ of Θ, and

A2) for each $\vartheta \in \Theta$, there exists $t = t(\vartheta) > 0$ such that $E_\vartheta[\sup_{\vartheta_1\in\Theta} P(v|\vartheta_1)/P(v|\vartheta)]^t < \infty$, and

A3) if $\vartheta, \vartheta_1 \in \Theta$ and $\vartheta \neq \vartheta_1$, then $\int 1_{P(v|\vartheta)\neq P(v|\vartheta_1)} dv > 0$ (here dv is the counting measure),

then there exists $0 < R < 1$, such that

$$P(\|\hat{\theta}^{(N)} - \theta\| \geq \epsilon|\theta) \leq R^N \tag{8}$$

for $\theta \in \Theta$, $\epsilon > 0$, and all sufficiently large N. □

Proof of Proposition 3.

Following the same reasoning as in Proposition 1(a), $\hat{\theta}^{(N)} \xrightarrow{P} \theta$.

The compactness of Θ satisfies A1 according to [5], the finiteness of $|\{0,1\}^J|$ satisfies A2, so that t can be any positive real number, and the identifiability of $P(v|\vartheta)$ satisfies A3. We let $\delta := \min\{|\theta_d| : \theta_d \neq 0\}$. From Theorem 3, for sufficiently large N,

$$P(\|\hat{\theta}^{(N)} - \theta\| \geq \delta/2) \leq R^N, 0 < R < 1. \tag{9}$$

Let $\rho = -\log R > 0$, this becomes

$$P(\|\hat{\theta}^{(N)} - \theta\| < \delta/2) \geq 1 - e^{-\rho N}. \tag{10}$$

We note that if $\|\hat{\theta}_N - \theta\| < \delta/2$, the order of coefficients w_{ij} are preserved for elimination. Specifically, if $\theta_d = 0$, $\theta_{d'} \neq 0$, then $|\hat{\theta}_d^{(N)}| < \delta/2$, $|\hat{\theta}_{d'}^{(N)}| > |\theta_{d'}| - |\hat{\theta}_{d'}^{(N)} - \theta_{d'}| > \delta/2$. Thus $|\hat{\theta}_{d'}^{(N)}| > |\hat{\theta}_d^{(N)}|$. Given $p \geq p_0$, we immediately have $A \subset \hat{A}^{(N)}$ given $\|\hat{\theta}_N - \theta\| < \delta/2$.

For (a), let $\mathcal{A} \subset \mathcal{F} := \{1, \ldots, D\}$, and $\hat{\theta}_{\mathcal{A}} = \arg\max_{\vartheta_{\mathcal{F}\setminus\mathcal{A}}=0} l^{(N)}(\vartheta)$ denote the maximum likelihood estimate of θ with components limited to set $\mathcal{A}$. For any $\mathcal{A}$ such that $A \subset \mathcal{A} \subset \mathcal{F}$ and $\epsilon > 0$, we have $\lim_{N\to\infty} P(\|\hat{\theta}_{\mathcal{A}}^{(N)} - \theta\| \geq \epsilon) = 0$ as in Proposition 1(a) . Because $|\{\mathcal{A}\}|$ is finite,

$$\lim_{N\to\infty} \sup_{A\subset\mathcal{A}} P(\|\hat{\theta}_{\mathcal{A}}^{(N)} - \theta\| \geq \epsilon) = 0, \tag{11}$$

and hence,

$$\lim_{N\to\infty} P(\|\tilde{\theta}^{(N)} - \theta\| \geq \epsilon) \leq \lim_{N\to\infty} P(A \not\subset \hat{A}^{(N)}) + \lim_{N\to\infty} \sup_{A\subset\mathcal{A}} P(\|\hat{\theta}_{\mathcal{A}}^{(N)} - \theta\| \geq \epsilon) = 0. \tag{12}$$

□

Proof of Corollary 1.

Suppose for sufficiently large N, $P(A \subset \hat{A}^{(N),r}) \geq 1 - e^{-\rho n}$, where r denote the r-th elimination. Let $\hat{A}_{\mathcal{A}}^{(N)}$ denote the remaining components when they are already restricted to a set $\mathcal{A}$. Because $|\{\mathcal{A}\}|$ is finite, there exists $N_1 \in \mathbb{N}$ and $\rho' > 0$ such that for $N \geq N_1$, $\inf_{A\subset\mathcal{A}} P(A \subset \hat{A}_{\mathcal{A}}^{(N)}) \geq 1 - e^{-\rho' n}$. When $N \geq N_1$ and is sufficiently large,

$$\begin{aligned} P(A \subset \hat{A}^{(N),r+1}) &\geq P(A \subset \hat{A}^{(N),r}) P(A \subset \hat{A}^{(N),r+1} | A \subset \hat{A}^{(N),r}) \\ &\geq P(A \subset \hat{A}^{(N),r}) \inf_{A\subset\mathcal{A}} P(A \subset \hat{A}_{\mathcal{A}}^{(N)}) > 1 - e^{-\rho N} - e^{\rho' N}. \end{aligned} \tag{13}$$

Therefore for N large enough, $P(A \subset \hat{A}^{(N),r+1}) \geq 1 - e^{-\rho'' N}$, ρ'' being any positive number smaller than $\min\{\rho, \rho'\}$. By induction from each r to $r+1$, part (b) holds. Part (a) holds following the same reasoning as in the proof of Proposition 3(a). □

Proof of Proposition 4.

We consider a certain mask m for a Dropout/DropConnect RBM. For a Dropout RBM, the energy function can be written as

$$\mathcal{E}(v, h, m) = -b^T(m_v * v) - c^T(m_h * h) - (m_h * h)^T W^\star (m_v * v). \tag{14}$$

Note that we also drop visible nodes in this scenario in order to keep symmetry, so $m = (m_v, m_h)$. For a DropConnect RBM,

$$\mathcal{E}(v, h, m) = -b^T v - c^T h - h^T (m * W) v. \tag{15}$$

Suppose we add a hidden node $h^\star$ to the RBM, and denote the mask applied to $h^\star$ as $m^\star$. Then we have for a Dropout RBM

$$\begin{aligned}&\mathcal{E}(v,h,m,h^\star,m^\star)\\ &= -b^T(m_v * v) - c^T(m_h * h) - c^\star m^\star h^\star - (m_h * h)^T W(m_v * v) - (m^\star h^\star)\cdot W^\star(m_v * v),\end{aligned} \tag{16}$$

and for a DropConnect RBM,

$$\mathcal{E}(v,h,m,h^\star,m^\star) = -b^T v - c^T h - c^\star h^\star - h^T(m * W)v - h^\star \cdot (m^\star * W^\star)v. \tag{17}$$

Initializing with $c^\star = 0$, $W^\star = 0$, we have for both models

$$\mathcal{E}(v,h,m,h^\star,m^\star) = \mathcal{E}(v,h,m). \tag{18}$$

The initial log-likelihood after adding the hidden node is

$$\begin{aligned}l^{(N)}_{new}(v^{(1)},\dots,v^{(N)}|\hat{\theta},c^*,W^*) &= \sum_{n=1}^{N} E_{m,m^\star}\log P_{new}(v^{(n)}|\hat{\theta})\\ &= \sum_{n=1}^{N} E_{m,m^\star}\log\frac{\sum_{h,h^\star} e^{-\mathcal{E}(v^{(n)},h,m,h^\star,m^\star)}}{\sum_{h,h^\star,v} e^{-\mathcal{E}(v,h,m,h^\star,m^\star)}}\\ &= \sum_{n=1}^{N} E_{m,m^\star}\log\frac{\sum_{h,h^\star} e^{-\mathcal{E}(v^{(n)},h,m)}}{\sum_{h,h^\star,v} e^{-\mathcal{E}(v,h,m)}}\\ &= \sum_{n=1}^{N} E_{m}\log\frac{2\sum_{h} e^{-\mathcal{E}(v^{(n)},h,m)}}{2\sum_{h,v} e^{-\mathcal{E}(v,h,m)}} \qquad (h^\star\in\{0,1\})\\ &= l^{(N)}(v^{(1)},\dots,v^{(N)}|\hat{\theta}).\end{aligned} \tag{19}$$

With the CD-k algorithm, the likelihood does not decrease after adding a new hidden node. Specifically, $l^{(N)}_{new}(v^{(1)},\dots,v^{(N)}|\hat{\theta}^\star) \geq l^{(N)}_{new}(v^{(1)},\dots,v^{(N)}|\hat{\theta},c^*,W^*) = l^{(N)}(v^{(1)},\dots,v^{(N)}|\hat{\theta})$, if $\hat{\theta}^*$ is the optimal value after adding a new node.

For adding layers under Dropout and DropConnect, following [1, 3], we illustrate in details why

$$E_{m^*}[\log P_{RBM_{L+1}}(h^L|m^*)] = E_m[\log P_{DBN_L}(h^L|m)] \tag{20}$$

holds, and how adding a symmetric layer can improve the likelihood. Below, we let m^l denote the mask for the l-th layer.

$$\begin{aligned}\log P_{DBN_L}(h^L|m) &= \sum_{v,h^1,\dots,h^{L-1}} \log P_{RBM_1}(v|h^1,m^1)\cdots P_{RBM_L}(h^{L-1}|h^L,m^L)P_{RBM_L}(h^L,m^L)\\ &= \sum_{h^{L-1}} \log P_{RBM_L}(h^{L-1},h^L|m^L)\\ &= \sum_{h^{L+1}} \log P_{RBM_{L+1}}(h^L,h^{L+1}|m^*)\\ &= \log P_{RBM_{L+1}}(h^L|m^*),\end{aligned} \tag{21}$$

if m^L and m^* are symmetric. Because both of them are assumed to have constant dropping probabilities, from symmetry and taking an expectation over (21), we know that (20) holds. Therefore DBN_{L+1} has the same log-likelihood bound as DBN_L, and to improve the log-likelihood, we only need to improve $\sum_{h^L} E_{m,m^*}[P_{DBN_L}(h^L|v,m)\cdot\log P_{RBM_{L+1}}(h^L|m^*)]$ in equation (15) in the main paper. In training, we make the following approximation

$$\sum_{h^L} E_{m,m^*}[P_{DBN_L}(h^L|v,m)\cdot\log P_{RBM_{L+1}}(h^L|m^*)]$$

$$\doteq E_{m^*}\{\log P_{RBM_{L+1}}[E_m E_{DBN_L}(h^L|v,m)|m^*]\}, \tag{22}$$

where $E_m E_{DBN_L}(h^L|v,m)$ can be easily obtained from the greedy layer-wise training approach. Therefore the likelihood is improved. It is also immediate that adding layers with size $J \leq H^1 \leq H^2 \leq \cdots$ continually improves the likelihood, since we add $H^L - H^{L-2}$ hidden nodes to the L-th hidden layer after we add the L-th hidden layer ($L \geq 2$, $H^0 = J$). □

References.

[1] Bengio, Y. (2007) Learning deep architectures for AI. *Technical Report*, https://www.iro.umontreal.ca/~lisa/pointeurs/TR1312.pdf.

[2] Geyer, C. (1994) On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22**(4):1993-2010.

[3] Hinton, G., Osindero, S. & Teh, Y.-W. (2006) A fast learning algorithm for deep belief nets. *Neural Computation* **18**(7):1527-1554.

[4] Knight, K. & Fu, W. (2000) Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28**(5):1356-1378.

[5] Shen, X. (2001) On Bahadur efficiency and maximum likelihood estimation. *Statistica Sinica* **11**(2):479-498.

[6] Van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

[7] Zou, H. (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**: 1418-1429.