Recovery of simultaneous low rank and two-way sparse coefficient matrices

Mladen Kolar (mkolar@chicagobooth.edu)

Collaborators

Ming Yu (Chicago Booth)

Varun Gupta (Chicago Booth)

Zhaoran Wang (Northwestern)

Motivation — Multi-task learning

Learning multiple related tasks leads to better statistical performance compared to learning the tasks separately.

We consider the following linear regression multi-task learning setting

$$Y = X\Theta^* + E,$$

where

- $Y \in \mathbb{R}^{n \times k}$ is a matrix of responses
- $X \in \mathbb{R}^{n \times p}$ is a matrix of predictors
- $\Theta^* \in \mathbb{R}^{p \times k}$ is an unknown parameter matrix
- $E \in \mathbb{R}^{n \times k}$ is an error matrix with i.i.d. mean zero and variance σ^2 entries

Relatedness of tasks is modeled through structural assumptions on the matrix Θ^*.

Motivation — Multi-task learning

In a high-dimensional setting, with large number of variables, it is common to assume that there are a few variables predictive of all tasks, while others are not predictive

- Turlach et al. (2005), Obozinski et al. (2011), Lounici et al. (2011), Kolar et al. (2011), Wang et al. (2016b)

$$\widehat{\Theta} = \arg\min \frac{1}{2n} \|Y - X\Theta\|_F^2 + \sum_{j \in [p]} \mathrm{pen}(\Theta_{j\cdot})$$

where $\mathrm{pen}(\cdot)$ is usually ℓ_2 or ℓ_∞ norm.

Motivation — Multi-task learning

Another way to relate tasks is to assume that predictors lie in a shared lower dimensional subspace

- Ando and Zhang (2005), Amit et al. (2007), Yuan et al. (2007), Argyriou et al. (2008), Wang et al. (2016a)

That is, Θ^* is assumed to be a low rank matrix.

Bunea et al. (2011) show optimality for the following reduced rank estimator

$$\widehat{\Theta} = \arg\min \frac{1}{2n} \| Y - X\Theta \|_F^2 + \lambda \cdot \mathrm{rank}(\Theta),$$

which can be efficiently computed using SVD (Reinsel and Velu, 1998).

Motivation — Multi-task learning

More commonly, one uses a relaxation of the rank constraint.

$$\widehat{\Theta} = \arg\min \frac{1}{2n} \|Y - X\Theta\|_F^2 + \lambda \cdot \|\Theta\|_*,$$

where $\|\Theta\|_* = \sum_{j=1}^{\mathrm{rank}(\Theta)} \sigma_j(\Theta)$ is the nuclear norm.

See, for example, (Candès and Recht, 2009, Chandrasekaran et al. (2011), Koltchinskii et al. (2011), Harchaoui et al. (2012), Negahban and Wainwright (2011), ...)

Sparse reduced rank regression

In contemporary applications it is increasingly common that both the number of predictors and the number of tasks is large compared to the sample size.

- In a study of regulatory relationships between genome-wide measurements, where micro-RNA measurements are used to explain the gene expression levels, a small number of micro-RNAs regulate genes participating in few regulatory pathways (Ma et al., 2014a).

Θ^* is assumed to be both sparse and low rank.

- predictors can be combined into fewer latent features that drive the variation in the multiple response variables and are composed only of relevant predictor variables
- Bunea et al. (2012), Chen et al. (2012), Chen and Huang (2012), She (2017)

More applications

Sparse SVD

- Chen et al. (2012), Ma et al. (2014a), Yang et al. (2014), ...

Biclustering:

- Lee et al. (2010), Balakrishnan et al. (2011), Balakrishnan et al. (2017)

Optimization over sparse and low-rank matrices

We consider a statistical model with true parameter $\Theta^* \in \Omega$, where $\Omega \subset \mathbb{R}^{m_1 \times m_2}$ is a nonconvex set comprising of low rank matrices that are also row and/or column sparse,

$$\Omega = \Omega(r, s_1, s_2) = \{\Theta \mid \mathrm{rank}(\Theta) \leq r, \|\Theta\|_{2,0} \leq s_1, \|\Theta^\top\|_{2,0} \leq s_2\},$$

with $\|\Theta\|_{2,0}$ is the number of non-zero rows of Θ.

To estimate Θ^*, we minimize an empirical loss function

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} f(\Theta)$$

over the set Ω.

Not clear how to do a convex relaxation for the set Ω.

Optimization over sparse and low-rank matrices

We write $\Theta = UV^\top$ with $U \in \mathbb{R}^{m_1 \times r}, V \in \mathbb{R}^{m_2 \times r}$ and consider the following optimization problem

$$(\widehat{U}, \widehat{V}) \in \arg \min_{U \in \mathcal{U}, V \in \mathcal{V}} f(U, V),$$

where

$$\begin{aligned} \mathcal{U} &= \mathcal{U}(s_1) = \left\{ U : \|U\|_{2,0} \leq s_1 \right\}, \\ \mathcal{V} &= \mathcal{V}(s_2) = \left\{ V : \|V\|_{2,0} \leq s_2 \right\}. \end{aligned}$$

- $\widehat{U}$ and $\widehat{V}$ are only unique up to rotation: $(\widehat{U}R, \widehat{V}R)$ is also a solution for any orthogonal matrix R.

Burer-Monteiro factorization for low rank matrices

Low-rank Matrix Recovery

$$\min_{\Theta \in \mathbb{R}^{m_1 \times m_2}} f(\Theta) \quad \text{subject to } \mathrm{rank}(\Theta) \leq r,$$

Convex relaxation

$$\min_{\Theta \in \mathbb{R}^{m_1 \times m_2}} f(\Theta) + \lambda \|\Theta\|_*.$$

Nonconvex approach

- Write $\Theta = UV^\top$ with $U \in \mathbb{R}^{m_1 \times r}$ and $V \in \mathbb{R}^{m_2 \times r}$ and minimize

$$\min_{U,V} f(U, V)$$

Burer-Monteiro factorization in practice

More efficient than solving convex relaxation

Good performance with good objective function and initialization

Nonconvex optimization in theory

- Keshavan et al. (2010), Jain et al. (2013), Hardt (2014), Hardt et al. (2014), Zhao et al. (2015), Zheng and Lafferty (2015), Bhojanapalli et al. (2016), Zhu et al. (2017), Ge et al. (2016), Li et al. (2017)

Rotation issue

We use the penalty function $g(U, V)$ defined as

$$g(U, V) = \frac{1}{4}\left\|U^\top U - V^\top V\right\|_F^2,$$

which forces U and V to be balanced (Zheng and Lafferty, 2015).

We now consider the following problem

$$(\widehat{U}, \widehat{V}) \in \arg\min_{U \in \mathcal{U}, V \in \mathcal{V}} f(U, V) + g(U, V).$$

The solution will be the same as the previous problem.

Rotation issue — Measuring convergence

Subspace distance.

- Denote $Z = [U; V]$, $Z^* = [U^*; V^*]$ with $U^*(V^*)^\top = \Theta^*$ and $U^*(U^*)^\top = V^*(V^*)^\top$, we define the subspace distance as:

$$d(Z, Z^*) = \min_{R \in \mathbb{Q}^r} \left\{ \|U - U^* R\|_F + \|V - V^* R\|_F \right\},$$

where $\mathbb{Q}^r$ denotes the set of r-by-r orthogonal matrixes.

We will show that $d(Z^t, Z^*)$ converges linearly up to statistical error.

Algorithm

Algorithm 1 Gradient Descent with Hard Thresholding (GDT)

Input: Initial estimate $\widetilde{\Theta}$
Parameters: Step size η, Rank r, Sparsity s_1, s_2, Number of iterations T
$(\widetilde{U}, \widetilde{\Sigma}, \widetilde{V})$ = rank r SVD of $\widetilde{\Theta}$
$U^0 = \text{Hard}(\widetilde{U}(\widetilde{\Sigma})^{\frac{1}{2}}, s_1)$, $V^0 = \text{Hard}(\widetilde{V}(\widetilde{\Sigma})^{\frac{1}{2}}, s_2)$
for $t = 1$ **to** T **do**
    $V^{t+0.5} = V^t - \eta \nabla_V f(U^t, V^t) - \eta \nabla_V g(U^t, V^t)$,
    $V^{t+1} = \text{Hard}(V^{t+0.5}, s_2)$
    $U^{t+0.5} = U^t - \eta \nabla_U f(U^t, V^t) - \eta \nabla_U g(U^t, V^t)$,
    $U^{t+1} = \text{Hard}(U^{t+0.5}, s_1)$
end for
Output: $\Theta^T = U^T (V^T)^\top$

Hyperparameters

Rank r

- Using ideas from Bunea et al. (2011).

Sparsity levels s_1, s_2

- Use $s_1 = c \cdot s_1^*$ and $s_2 = c \cdot s_2^*$ with some $c > 1$.
- Information criteria, such as She (2017).
- Not very sensitive to the choice of c.

Our algorithm does not require tuning parameters that need to be selected carefully other than the rank, which is required for most of the methods.

Assumptions

Restricted Strong Convexity and Smoothness (RSC/RSS)

There exist universal constants μ and L such that

$$\frac{\mu}{2}\|\Theta_2-\Theta_1\|_F^2 \leq f(\Theta_2)-f(\Theta_1)-\langle \nabla f(\Theta_1), \Theta_2-\Theta_1\rangle \leq \frac{L}{2}\|\Theta_2-\Theta_1\|_F^2 \quad (1)$$

for all $\Theta_1, \Theta_2 \in \Omega(2r, \widetilde{s}_1, \widetilde{s}_2)$ where $\widetilde{s}_1 = (2c+1)s_1^*$ and $\widetilde{s}_2 = (2c+1)s_2^*$.

Assumptions

Initialization (I)

Define $\mu_{\min} = \frac{1}{8}\min\{1, \frac{\mu L}{\mu+L}\}$ and

$$I_0 = \frac{4}{5}\mu_{\min}\sigma_r(\Theta^*) \cdot \min\left\{\frac{1}{\mu+L}, 2\right\}.$$

We require

$$\|\Theta^0 - \Theta^*\|_F \leq \frac{1}{5}\min\left\{\sigma_r(\Theta^*), \frac{I_0}{\xi}\sqrt{\sigma_r(\Theta^*)}\right\},$$

where $\xi^2 = 1 + \frac{2}{\sqrt{c-1}}$.

Assumptions

We define the notion of the statistical error,

$$e_{\mathsf{stat}} = \sup_{\substack{\Delta \in \Omega(2r, \widetilde{s_1}, \widetilde{s_2}) \\ \|\Delta\|_F \leq 1}} \langle \nabla f(\Theta^*), \Delta \rangle.$$

Step Size Selection: We choose the step size η to satisfy

$$\eta \leq \frac{1}{16\|Z_0\|_2^2} \cdot \min\left\{\frac{1}{2(\mu + L)}, 1\right\}.$$

Furthermore, we require η and c to satisfy

$$\beta = \xi^2 \left(1 - \eta \cdot \frac{2}{5}\mu_{\min}\sigma_r(\Theta^*)\right) < 1,$$

and

$$e_{\mathrm{stat}}^2 \leq \frac{1-\beta}{\xi^2 \eta} \cdot \frac{L\mu}{L+\mu} \cdot I_0^2.$$

Key Lemma

Suppose the conditions **(RSC/RSS)**, **(I)** are satisfied. Assume that the point $Z = \begin{bmatrix} U \\ V \end{bmatrix}$ satisfies $d(Z, Z^*) \leq l_0$. Let (U^+, V^+) denote the next iterate obtained with GDT with the step size η satisfying

$$\eta \leq \frac{1}{8\|Z\|_2^2} \cdot \min\left\{\frac{1}{2(\mu + L)}, 1\right\}.$$

Then we have

$$d^2(Z^+, Z^*) \leq \xi^2 \left[\left(1 - \eta \cdot \frac{2}{5}\mu_{\min}\sigma_r(\Theta^*)\right) \cdot d^2(Z, Z^*) + \eta \cdot \frac{L + \mu}{L \cdot \mu} \cdot e_{\mathrm{stat}}^2\right],$$

where $\xi^2 = 1 + \frac{2}{\sqrt{c-1}}$.

Main Result

Suppose the conditions **(RSC/RSS)**, **(I)** are satisfied and the step size η satisfies the conditions stated before. Then after T iterations of GDT, we have

$$d^2(Z^T, Z^*) \leq \beta^T \cdot d^2(Z^0, Z^*) + \frac{\xi^2 \eta}{1-\beta} \cdot \frac{L+\mu}{L \cdot \mu} \cdot e_{\mathrm{stat}}^2.$$

Furthermore, for $\Theta^T = U^T (V^T)^\top$ we have

$$\|\Theta^T - \Theta^*\|_F^2 \leq 4\sigma_1(\Theta^*) \cdot \left[\beta^T \cdot d^2(Z^0, Z^*) + \frac{\xi^2 \eta}{1-\beta} \cdot \frac{L+\mu}{L \cdot \mu} \cdot e_{\mathrm{stat}}^2\right].$$

- Our analysis also works for optimization problem without statistical model, where we replace true values U^*, V^* with global minimum $\widehat{U}, \widehat{V}$. If we further assume no sparsity, the statistical error is 0.

Application to Multi-task Learning

Recall that we are interested in a multi-task learning problem

$$Y = X\Theta^* + E,$$

where

- $Y \in \mathbb{R}^{n\times k}$ is a matrix of responses
- $X \in \mathbb{R}^{n\times p}$ is a matrix of predictors
- $\Theta^* \in \mathbb{R}^{p\times k}$ is an unknown parameter matrix
- $E \in \mathbb{R}^{n\times k}$ is an error matrix with i.i.d. mean zero and variance σ^2 entries

The objective function is

$$f(U, V) = \frac{1}{2n}\|Y - XUV^\top\|_F^2$$

with $U \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{R}^{k\times r}$ with $U \in \mathcal{U}(s_1)$ and $V \in \mathcal{U}(s_2)$.

Application to Multi-task Learning

We assume X satisfies the Restricted Eigenvalue (RE) condition (Negahban et al., 2012) for some constant $\underline{\kappa}(s_1)$ and $\bar{\kappa}(s_1)$

$$\underline{\kappa}(s_1) \cdot \|\theta\|_2^2 \leq \frac{1}{n}\|X\theta\|_2^2 \leq \bar{\kappa}(s_1) \cdot \|\theta\|_2^2 \quad \text{for all } \|\theta\|_0 \leq s_1,$$

which implies that the **(RSC/RSS)** condition is satisfied.

Initialization is done using a lasso estimator. The condition **(I)** is effectively a requirement on the sample size.

Application to Multi-task Learning

Suppose all the conditions are satisfied, for all

$$T \geq C_1 \log \left[\frac{n}{(s_1^* + s_2^*)\big(r + \log(p \vee k)\big)} \right],$$

with probability at least $1 - (p \wedge k)^{-1}$, we have

$$\|\Theta^T - \Theta^*\|_F \leq C\sigma \sqrt{\frac{(s_1^* + s_2^*)\big(r + \log(p \vee k)\big)}{n}}$$

for some constant C_1 and C.

Application to Multi-task Learning

We compare the error rate

$$\sigma\sqrt{\frac{1}{n}(s_1^* + s_2^*)\big(r + \log(p \vee k)\big)}$$

with the minimax rate established in (Ma et al., 2014b):

$$\sigma\sqrt{\frac{1}{n}\left[(s_1^* + s_2^*)r + s_1^* \log\frac{ep}{s_1^*} + s_2^* \log\frac{ek}{s_2^*}\right]}$$

They match up to a $\log(p \vee k)$ factor. When r is comparable to $\log(p \vee k)$ they match up to a constant multiplier.

For large enough T, GDT algorithm attains near optimal rate.

Application to Multi-task Learning

If we consider row sparsity only, then we have $s_2^* = k$ and

$$\|\Theta^T - \Theta^*\|_F \leq C\sigma\sqrt{\frac{kr + s_1^*(r + \log p)}{n}}.$$

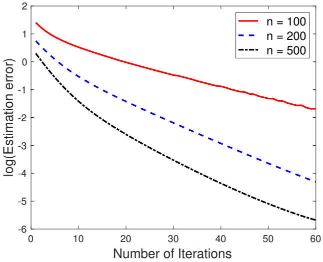
This gives prediction error

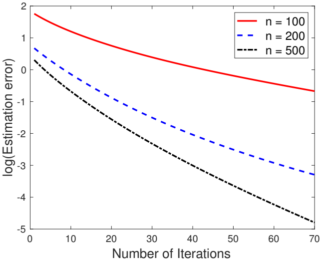
$$\|X\Theta^T - X\Theta^*\|_F^2 \leq C\sigma^2\Big(kr + s_1^*(r + \log p)\Big).$$

GDT error matches the prediction error $(k + s_1^* - r)r + s_1^* \log p$ provided in (She, 2017), as long as $k \geq Cr$ which is typically satisfied.

Experiment

Linear convergence





Simulation

Comparison with other methods

- Double Projected Penalization (DPP) — Ma et al. (2014b)
- thresholding SVD method (TSVD) — Ma et al. (2014a)
- exclusive extraction algorithm (EEA) — Chen et al. (2012)
- RCGL and JRRS — Bunea et al. (2012)
- standard Multitask learning method (MTL, with $L_{2,1}$ penalty)

Setup: $n = 50, p = 100, k = 50, r = 8, s_1^* = s_2^* = 10.$

For the methods that rely on a tuning parameter λ, we generate an independent validation set to select the "best" λ.

For our method, we use $s_1 = 2s_1^*$ and $s_2 = 2s_2^*$.

Simulation

Table 1: Row sparse

	Estimation error	Prediction error	$\|$Row support$\|$
GDT	0.0488 ± 0.0103	1.1043 ± 0.0144	20 ± 0
DPP	0.0588 ± 0.0148	1.1079 ± 0.0155	48.96 ± 8.29
TSVD	0.3169 ± 0.1351	2.4158 ± 0.9899	25.62 ± 8.03
EEA	0.3053 ± 0.0998	1.2349 ± 0.0362	84.28 ± 6.70
RCGL	0.0591 ± 0.0148	1.1101 ± 0.0168	49.60 ± 10.62
JRRS	0.0877 ± 0.0227	1.1857 ± 0.0214	12.26 ± 2.02
MTL	0.0904 ± 0.0243	1.1753 ± 0.0204	73.40 ± 2.67

Simulation

Table 2: Row sparse and column sparse

	Estimation error	Prediction error	Row support	Column support
GDT	0.087 ± 0.023	1.062 ± 0.014	20 ± 0	20 ± 0
DPP	0.098 ± 0.028	1.044 ± 0.014	51.3 ± 13.9	10.2 ± 0.5
TSVD	0.335 ± 0.105	1.760 ± 0.341	28.6 ± 7.2	30.9 ± 8.5
EEA	0.260 ± 0.115	1.102 ± 0.022	64.4 ± 9.8	12.1 ± 2.7
RCGL	0.121 ± 0.032	1.107 ± 0.017	42.0 ± 7.9	50 ± 0
JRRS	0.168 ± 0.041	1.161 ± 0.017	13.9 ± 4.6	50 ± 0
MTL	0.183 ± 0.049	1.165 ± 0.016	73.5 ± 3.1	50 ± 0

Simulation

- Increase n, p, s_1^*, s_2^* by a factor of ζ
- Increase k, r by a factor of $\lfloor\sqrt{\zeta}\rfloor$

Table 3: Running time comparison (in seconds)

	$\zeta = 1$	$\zeta = 5$	$\zeta = 10$	$\zeta = 20$	$\zeta = 50$	$\zeta = 100$
GDT	0.11	0.20	0.51	2.14	29.3	235.8
DPP	0.19	0.61	3.18	17.22	315.4	2489
TSVD	0.07	1.09	6.32	37.8	543	6075
EEA	0.50	35.6	256	$>$2h	$>$2h	$>$2h
RCGL	0.18	1.02	7.15	36.4	657.4	$>$2h
JRRS	0.19	0.82	6.36	30.0	610.2	$>$2h
MTL	0.18	3.12	30.92	184.3	$>$2h	$>$2h

In vivo Calcium Imaging Data

When a neuron fires an electrical action potential, calcium will enter the cell and then its fluorescent properties.

By recording the movies of this dynamic it allows us to identify the spiking activity from large populations of neurons.

Spatiotemporal model introduced by (Pnevmatikakis et al. (2014))

In vivo Calcium Imaging Data

Observation field: $k = \ell_1 \times \ell_2$ pixels

The field contains a total number of (possibly overlapping) r neurons

Let $\boldsymbol{c}_i$ denote the calcium activity for each neuron i, it follows AR(1) model:

$$c_i(t) = \gamma c_i(t-1) + s_i(t),$$

where $s_i(t)$ is the number of spikes that neuron i fired at time t and $\gamma = 1 - 1/(\text{frame rate})$.

Let $\boldsymbol{a}_i$ denote the spatial footprint vector for neuron i, our observation at each time step t is

$$\boldsymbol{y}(t) = \sum_{i=1}^{K} \boldsymbol{a}_i c_i(t) + \epsilon_t.$$

In vivo Calcium Imaging Data

In matrix form we can rewrite as

$$S = GC$$
$$Y = CA + E$$

with

$$G = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\gamma & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & -\gamma & 1 \end{pmatrix}.$$

Here $C \in \mathbb{R}^{T\times r}, G \in \mathbb{R}^{T\times T}, S \in \mathbb{R}^{T\times r}, Y \in \mathbb{R}^{T\times k}$ and $A \in \mathbb{R}^{r\times k}$.

In vivo Calcium Imaging Data

Combine them together we obtain

$$Y = G^{-1}SA + E = X\Theta^* + E$$

where $X = G^{-1}$ is observed and $\Theta^* = SA$ is the coefficient matrix.

A should be row sparse since the area for neurons in the monitored area is small.

S should be column sparse since neurons do not fire very frequently.

Θ^* is low rank by construction since the number of neurons are usually small.

In vivo Calcium Imaging Data

Multi-task learning problem with simultaneous row-sparse, column-sparse and low rank coefficient matrix where $n = p = T$ and $k = \ell_1 \times \ell_2$.

The dataset is a movie with 559 frames (acquired at approximately 8.64 frames/sec), where each frame is 135×131 pixels.

We have $n = p = 559$ and $k = 135 \times 131 = 17,685$.

We use $r = 50$, more conservative than the estimator given by (Bunea et al., 2011) and we set $s_1 = 100$ row sparsity and $s_2 = 3000$ column sparsity.

In vivo Calcium Imaging Data

Figure 1: Manually selected top 5 labeled regions

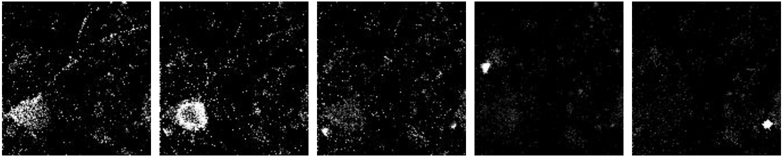


Figure 2: Corresponding signals estimated by our GDT algorithm

Conclusion

Nonconvex optimization on simultaneous low rank and two-way sparse coefficient matrix

GDT algorithm: alternating gradient descent with hard thresholding converges linearly to statistical error

For multi-task learning, statistical error is near optimal compared to the minimax rate

Better estimation accuracy and much faster running speed

References

Amit, Y., Fink, M., Srebro, N., and Ullman, S. (2007). Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24. ACM.

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Mach. Learn.*, 73(3):243–272.

Balakrishnan, S., Kolar, M., Rinaldo, A., and Singh, A. (2017). Recovering block-structured activations using compressive measurements. *Electron. J. Statist.*, 11(1):2647–2678.

Balakrishnan, S., Kolar, M., Rinaldo, A., Singh, A., and Wasserman, L. (2011). Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, volume 4.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016). Global optimality of local search for low rank matrix recovery. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 3873–3881. Curran Associates, Inc.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309.

Bunea, F., She, Y., and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, 40(5):2359–2388.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596.

Chen, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(2):203–221.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Ge, R., Lee, J. D., and Ma, T. (2016). Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981.

Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., and Malick, J. (2012). Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Hardt, M. (2014). Understanding alternating minimization for matrix completion. In *55th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2014*, pages 651–660. IEEE Computer Soc., Los Alamitos, CA.

Hardt, M., Meka, R., Raghavendra, P., and Weitz, B. (2014). Computational limits for matrix completion. In Balcan, M., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 703–725. JMLR.org.

Jain, P., Netrapalli, P., and Sanghavi, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *STOC'13—Proceedings of the 2013 ACM Symposium on Theory of Computing*, pages 665–674. ACM, New York.

Keshavan, R. H., Montanari, A., and Oh, S. (2010). Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998.

Kolar, M., Lafferty, J. D., and Wasserman, L. A. (2011). Union support recovery in multi-task learning. *J. Mach. Learn. Res.*, 12:2415–2435.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.*, 39(5):2302–2329.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095.

Li, Q., Zhu, Z., and Tang, G. (2017). Geometry of factored nuclear norm regularization. *Technical report*.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. A. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Stat.*, 39:2164–204.

Ma, X., Xiao, L., and Wong, W. H. (2014a). Learning regulatory programs by threshold SVD regression. *Proceedings of the National Academy of Sciences*, 111(44):15675–15680.

Ma, Z., Ma, Z., and Sun, T. (2014b). Adaptive estimation in two-way sparse reduced-rank regression. *Technical report*.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Stat.*, 39(2):1069–1097.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers. *Stat. Sci.*, 27(4):538–557.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Stat.*, 39(1):1–47.

Pnevmatikakis, E. A., Gao, Y., Soudry, D., Pfau, D., Lacefield, C., Poskanzer, K., Bruno, R., Yuste, R., and Paninski, L. (2014). A structured matrix factorization framework for large scale calcium imaging data analysis. *Technical report*.

She, Y. (2017). Selective factor extraction in high dimensions. *Biometrika*, 104(1):97–110.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Wang, J., Kolar, M., and Srebro, N. (2016a). Distributed multi-task learning with shared representation. *Technical report*.

Wang, J., Kolar, M., and Srerbo, N. (2016b). Distributed multi-task learning. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 751–760, Cadiz, Spain. PMLR.

Yang, D., Ma, Z., and Buja, A. (2014). A sparse singular value decomposition method for high-dimensional data. *J. Comput. Graph. Statist.*, 23(4):923–942.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. B*, 69(3):329–346.

Zhao, T., Wang, Z., and Liu, H. (2015). Nonconvex low rank matrix factorization via inexact first order oracle. *Advances in Neural Information Processing Systems*.

Zheng, Q. and Lafferty, J. D. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 109–117. Curran Associates, Inc.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2017). The global optimization geometry of nonsymmetric matrix factorization and sensing. *Technical report*.