Spectral k-Support Norm Regularization

Andrew McDonald

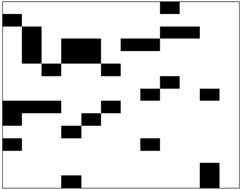
Department of Computer Science, UCL

(Joint work with **Massimiliano Pontil** and **Dimitris Stamos**)

25 March, 2015

Problem: Matrix Completion

Goal: Recover a matrix from a subset of measurements.



e.g. random sample of 10% of entries

Applications

- Collaborative filtering: predict interests of a user from preferences from many users (e.g. Netflix Problem)
- Triangulation of distances from an incomplete network (e.g. wireless network)
- Multitask learning: leverage commonalities between multiple learning tasks
- ...

Problem Statement

Given a subset Ω of observations of a matrix X, estimate the missing entries.

(i) ill-posed problem $\rightarrow$ assume X is low rank, use regularizer to encourage low rank structure

(ii) regularization with *rank* operator is NP hard $\rightarrow$ use convex approximation, e.g. *trace norm* (sum of singular values)

$$\min_{W} \|\Omega(W) - \Omega(X)\|_F^2 + \lambda \|W\|_{tr}$$

Trace Norm Regularization

- Trace norm is the tightest convex relaxation of *rank* operator on the spectral norm unit ball. [Fazel, Hindi & Boyd 2001]
- Optimization can be solved efficiently using proximal gradient methods.
- Can be shown that this method finds true underlying matrix with high probability.

Goal: can we improve on the performance using other regularizers?

The Vector k-Support Norm

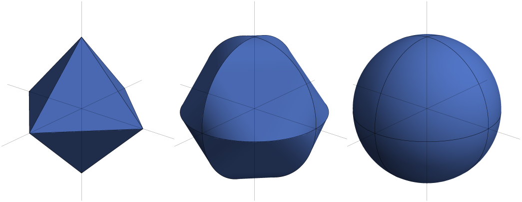
- The *k-support norm* is a regularizer used in sparse vector estimation problems. [Argyriou, Foygel & Srebro 2012]
- For $k \in \{1, \ldots, d\}$, unit ball is :

$$\mathrm{co}\{w : \mathrm{card}(w) \leq k, \|w\|_2 \leq 1\}.$$

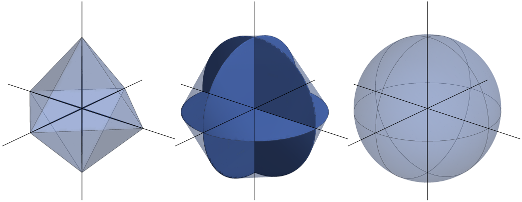
- Includes $\|\cdot\|_1$ ($k = 1$) and $\|\cdot\|_2$ ($k = d$).
- Dual norm is the ℓ_2-norm of the largest k components of a vector.

Vector k-Support Unit Balls

- unit balls in $\mathbb{R}^3$ ($k = 1, 2, 3$)



- convex hull interpretation



The Spectral k-Support Norm

Extend the k-support norm to matrices.

- The k-support norm is a symmetric gauge function: induces the *spectral k-support norm* [von Neumann 1937]

$$\|W\|_{(k)} = \|(\sigma_1(W), \ldots, \sigma_d(W))\|_{(k)}$$

- Unit ball is given by

$$\mathrm{co}\{W : \mathrm{rank}(W) \leq k, \|W\|_F \leq 1\}.$$

- Includes $\|\cdot\|_{tr}$ $(k = 1)$, and $\|\cdot\|_F$ $(k = d)$.

[McDonald, Pontil & Stamos 2014]

Optimization

- The k-support norm can be written as

$$\|w\|_{(k)} = \inf_{\theta \in \Theta} \sqrt{\sum_{i=1}^{d} \frac{w_i^2}{\theta_i}}, \quad \Theta = \{0 < \theta_i \leq 1, \sum_i \theta_i \leq k\}$$

- Coordinate-wise separable using Lagrange multipliers.
- The norm can be computed in $\mathcal{O}(d \log d)$ time as

$$\|w\|_{(k)}^2 = \|w_{[1:r]}^{\downarrow}\|_2^2 + \frac{1}{k-r}\|w_{(r:d]}^{\downarrow}\|_1^2.$$

- Similar computation for proximity operator of squared norm: can use proximal gradient methods to solve optimization.
- Matrix case follows using SVD.

Extension: The (k, p)-Support Norm

Fit the curvature of the underlying model.

- For $p \in [1, \infty]$ define the vector (k, p)-support norm by its unit ball

$$\mathrm{co}\{w : \mathrm{card}(w) \leq k, \|w\|_p \leq 1\}.$$

- The dual norm is the ℓ_q-norm of the k largest components ($\frac{1}{p} + \frac{1}{q} = 1$).

(Work in progress.)

The Spectral (k, p)-Support Norm

Fit the curvature of the underlying spectrum.

- For $p \in [1, \infty]$ the spectral (k, p)-support unit ball is defined in terms of the Schatten p-norm

$$\mathrm{co}\{W : \mathrm{rank}(W) \leq k, \|W\|_p \leq 1\}.$$

- Von Neumann again: $\|W\|_{(k,p)} = \|\sigma(W)\|_{(k,p)}$.

Optimization

- For $p \in (1, \infty)$ the (k, p)-support norm can be computed as

$$\|w\|_{(k,p)}^p = \|w^{\downarrow}_{[1:r]}\|_p^p + \frac{1}{(k-r)^{p/q}} \|w^{\downarrow}_{(r:d]}\|_1^p.$$

- For $p = 1$ we recover the ℓ_1 norm for all k, and for $p = \infty$ we have

$$\|w\|_{(k,\infty)} = \max\left(\|w\|_\infty, \frac{1}{k}\|w\|_1\right).$$

- For $p \in (1, \infty)$, we solve the constrained problem

$$\operatorname*{argmin}_s \left\{\langle s, \nabla \ell(w)\rangle : \|s\|_{(k,p)} \leq \alpha\right\}.$$

- For $p = \infty$ we can compute the projection onto the unit ball: can use proximal gradient methods.

Experiments: Matrix Completion

Benchmark datasets: MovieLens (movies), Jester (jokes)

dataset	norm	test error	k	p
MovieLens 100k	trace	0.2017	-	-
	k-support	0.1990	1.87	-
	(k, p)-support	**0.1988**	2.00	1.16
Jester 1	trace	0.1752	-	-
	k-support	0.1739	6.38	-
	(k, p)-support	**0.1731**	2.00	6.50
Jester 3	trace	0.1959	-	-
	k-support	0.1942	2.13	-
	(k, p)-support	**0.1932**	3.00	1.14

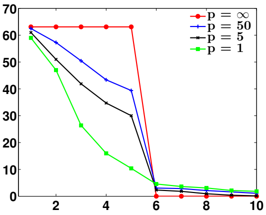
Note: $k = 1$ is trace norm, $p = 2$ is spectral k-support norm.

Role of p in (k, p)-Support Norm

Spectral (k, p)-support norm:

- Intuition: for large p, the ℓ_p norm of a vector is increasingly dominated by the largest components.
- Regularization with larger values of p encourages matrices with flatter spectrum.

Spectrum of synthetic rank 5 matrix with different regularizers:



Extension: Connection to the Cluster Norm

- Using the infimum formulation

$$\|w\|_{box} = \inf_{\theta \in \Theta} \sqrt{\sum_{i=1}^{d} \frac{w_i^2}{\theta_i}}, \quad \Theta = \{a < \theta_i \leq b, \sum_i \theta_i \leq c\}.$$

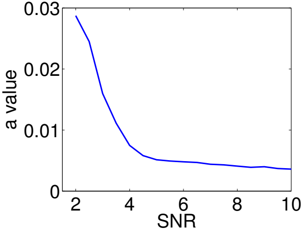
- *Box norm* is a perturbation of the k-support norm ($k = \frac{d-ca}{b-a}$)

$$\|w\|_{box}^2 = \min_{u,v} \left\{ \frac{1}{a}\|u\|_2^2 + \frac{1}{b-a}\|v\|_{(k)}^2 \right\}$$

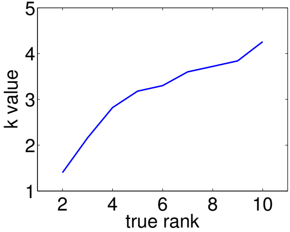
- Matrix case: we recover the *cluster norm* [Jacob, Bach, & Vert] used in multitask learning.

Role of a, c in Box Norm

Simulated datasets:



Low signal/high noise: high a.



High rank of underlying matrix: high k.

Further Work

- Statistical bounds on the performance of the norms: various results known [Chatterjee, Chen & Banerjee 2014, Maurer & Pontil 2012, Richard, Obozinksi & Vert 2014]
- Infimum formulation of (k, p)-support norm: known for $p \in [1, 2]$, unclear for $p \in (2, \infty]$.
- Study the family of norms for a general choice of the parameter set Θ. [Micchelli, Morales & Pontil 2013]

Conclusion

- Spectral k-support norm as regularizer for low rank matrix learning
- Spectral (k, p)-support norm allows us to learn curvature of the spectrum
- Box norm as perturbation of k-support norm
- Connection to multitask learning cluster norm

References

[1] A. Argyriou, R. Foygel, and N. Srebro.
Sparse prediction with the k-support norm.
In *Advances in Neural Information Processing Systems 25*, pages 1466–1474, 2012.

[2] S. Chatterjee, S. Chen, and A. Banerjee.
Generalized dantzig selector: Application to the k-support norm.
In *NIPS*, 2014.

[3] Maryam Fazel, Haitham Hindi, and Stephen P. Boyd.
A rank minimization heuristic with application to minimum orders system approximation.
Proceedings of the American Control Conference, 2001.

[4] L. Jacob, F. Bach, and J.-P. Vert.
Clustered multi-task learning: a convex formulation.
Advances in Neural Information Processing Systems (NIPS 21), 2009.

[5] A. Maurer and M. Pontil.
Structured sparsity and generalization.
The Journal of Machine Learning Research, 13:671–690, 2012.

[6] A. M. McDonald, M. Pontil, and D. Stamos.
Spectral k-support regularization.
In *Advances in Neural Information Processing Systems 27*, 2014.

[7] C. A. Micchelli, J. M. Morales, and M. Pontil.
Regularizers for structured sparsity.
Advances in Comp. Mathematics, 38:455–489, 2013.

[8] E. Richard, G. Obozinksi, and J.-P. Vert.
Tight convex relaxations for sparse matrix factorization.
In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[9] J. Von Neumann.
Some matrix-inequalities and metrization of matric-space.
Tomsk. Univ. Rev. Vol I, 1937.