Workshop on Nonsmooth Optimization and Applications NOPTA 8-12 April 2024 University of Antwerp, Belgium

Sparse phase retrieval and PCA: an optimal convex approach and practical nonconvex algorithm

Andrew D. McRae

Institute of Mathematics, EPFL

andrew.mcrae@epfl.ch

EPFL

Abstract

This poster presents novel analysis and algorithms for solving sparse phase retrieval and sparse principal component analysis (PCA) with nonsmooth convex lifted matrix formulations. The key innovation is a new atomic matrix norm that, when used as regularization, promotes low-rank matrices with sparse factors. We show that convex programs with this atomic norm as a regularizer provide near-optimal sample complexity and error rate guarantees for sparse phase retrieval and sparse PCA. Although we do not know efficient algorithms for the convex programs, for the phase retrieval case we carefully analyze the program and its dual and thereby derive a practical heuristic nonconvex algorithm. We show empirically that this nonconvex algorithm performs similarly to existing state-of-the-art sparse phase retrieval algorithms. Based on joint work with Justin Romberg and Mark Davenport, published in [1].

Statistical motivation and theory

General application: in a high-dimensional statistics setting, estimate a rank-1 matrix $B_* = u_* v_*^\top$ whose factors u_*, v_* are sparse (more generally, B_* could be low-rank with sparse factors).

Algorithmic goal: find a (convex) matrix norm that promotes this sparse and low-rank structure.

Challenge: *simultaneous* structure of B_*. B_* is not merely both sparse and low-rank; the factors of its low-rank decomposition are themselves sparse.

Low rank is often promoted with the nuclear norm:

$$\|B\|_* = \min \left\{ \sum \|u_k\|_2 \|v_k\|_2 : B = \sum u_k v_k^\top \right\}.$$

However, this does not account for sparsity. Matrix sparsity can be promoted with the (elementwise) ℓ_1 norm:

$$\|B\|_1 = \sum_{i,j} = |B_{ij}| = \min \left\{ \sum \|u_k\|_1 \|v_k\|_1 : B = \sum u_k v_k^\top \right\}.$$

However, the elementwise ℓ_1 norm does not account for low rank. Simply combining these norms does not work [2], [3].

New mixed atomic norm: we "mix" the nuclear and ℓ_1 norms into the following atomic norm:

$$\|B\|_{*,\gamma} := \min \left\{ \sum \theta_\gamma(u_k, v_k) : B = \sum u_k v_k^\top \right\}, \text{ where}$$
$$\theta_\gamma(u,v) := (\|u\|_2 + \gamma\|u\|_1)(\|v\|_2 + \gamma\|v\|_1).$$

$\gamma > 0$ is a parameter that controls the relative strength of the nuclear and ℓ_1 norm components.

Application: sparse phase retrieval. Suppose $\beta_* \in \mathbf{R}^d$ is sparse, and we observe

$$y_i = \langle x_i, \beta_* \rangle^2 + \xi_i = \langle X_i, B_* \rangle + \xi_i, i = 1, \dots, n$$

where $B_* := \beta_* \beta_*^\top$, and $X_i := x_i x_i^\top$. We estimate B_* by

$$\widehat{B} = \arg\min_{B \succeq 0} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle X_i, B \rangle)^2 + \lambda \|B\|_{*,\gamma}.$$

Theorem 1 (Simplified) *If β is s-sparse, the measurements $x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, the noise ξ_i is i.i.d. zero-mean and bounded, and λ, γ are chosen appropriately, then, if $n \gtrsim s \log \frac{d}{s}$,*

$$\|\widehat{B} - B_*\|_F^2 \lesssim \frac{s \log(d/s)}{n}.$$

The sample complexity and error rate are **optimal** (possibly modulo the log factor).

Application: sparse PCA. Suppose we observe $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbf{R}^{d \times d}$ has a sparse leading eigenvector v_1. We can estimate $P_* := v_1 v_1^\top$ with the following convex program:

$$\widehat{P} = \arg\min_{P \succeq 0} -\langle P, \widehat{\Sigma} \rangle + \lambda \|P\|_{*,\gamma} \text{ s.t. } \operatorname{tr}(P) \leq 1,$$

where $\widehat{\Sigma}$ is the empirical covariance.

Theorem 2 *If v_1 is s-sparse, and λ, γ are appropriately chosen, then, if $n \gtrsim s \log \frac{d}{s}$,*

$$\|\widehat{P} - P_*\|_F^2 \lesssim \frac{\sigma_1 \sigma_2}{(\sigma_1 - \sigma_2)^2} \frac{s \log(d/s)}{n},$$

where σ_1, σ_2 are the top two eigenvalues of Σ.

Again, the sample complexity and error rate are optimal modulo log factors.

General nonsmooth problem

Consider the following convex matrix program (the statistical estimators are special cases):

$$\min_{B \in \mathcal{C}} f(B) + \|B\|_{*,\gamma}, \tag{1}$$

where $\mathcal{C} \subseteq \mathbf{R}^{d_1 \times d_2}$ is convex, f is smooth and convex, $\gamma > 0$,

$$\|B\|_{*,\gamma} := \min \left\{ \sum \theta_\gamma(u_k, v_k) : B = \sum u_k v_k^\top \right\}, \text{ and}$$
$$\theta_\gamma(u,v) := (\|u\|_2 + \gamma\|u\|_1)(\|v\|_2 + \gamma\|v\|_1).$$

Although (1) is convex, it is unclear even how to evaluate (let alone optimize) the atomic norm component $\|B\|_{*,\gamma}$.

In fact, the sparse PCA performance achieved by Theorem 2 is widely believed to be impossible with polynomial-time estimators [4]. If this is true, then, **in general,** (1) **is computationally intractable.** Nevertheless, some specific instances may be tractable.

Nonconvex formulation: The convex problem (1) is equivalent to

$$\min \left\{ f(UV^\top) + \theta_\gamma(U, V) : r \geq 1, U \in \mathbf{R}^{d_1 \times r}, V \in \mathbf{R}^{d_2 \times r}, UV^\top \in \mathcal{C} \right\}, \tag{2}$$

where, if $U = [u_1, \dots, u_r], V = [v_1, \dots, v_r]$, we abbreviate $\theta_\gamma(U, V) = \sum_{k=1}^r \theta_\gamma(u_k, v_k)$.

In practice, we optimize U and V directly for fixed r and then update r. This approach is explored abstractly in [5]. In our case, the structure of θ_γ makes (2) amenable to **proximal methods**.

Algorithm for phase retrieval

We consider the following instance of (2):

$$\min_{\substack{r \\ U,V \in \mathbf{R}^{d \times r}}} \frac{1}{2n} \sum_{i=1}^n (y_i - \langle X_i, UV^\top \rangle)^2 + \lambda \theta_\gamma(U, V). \tag{3}$$

(Enforcing symmetry appears unnecessary in practice, and the asymmetric version is conveniently amenable to alternating minimization. Why this works is an open question.)

Optimality conditions: How do we check optimality of a candidate solution $B = UV^\top$? Let

$$Z := -\nabla f(B) = \frac{1}{n} \sum_{i=1}^n (y_i - \langle X_i, B \rangle_{\mathrm{HS}}) X_i.$$

B is a global optimum if and only if the following two conditions hold:

- **First-order criticality**: $\langle Z, u_k v_k^\top \rangle = \lambda \theta_\gamma(u_k, v_k)$ for $k = 1, \dots, r$.
- **Second-order criticality**: $\langle Z, uv^\top \rangle \leq \lambda \theta_\gamma(u, v)$ for **all** $u, v \in \mathbf{R}^d$. Any pair (u, v) violating this condition yields a descent direction if we add u and v (scaled sufficiently small) as additional columns to U and V.

This suggests the following meta-algorithm (adapted from [5]):

Algorithm 1: Sparse phase retrieval algorithm

```
input  : Data (X_i, y_i), i = 1, ..., n
output: Solution (r, U, V) to (3)
Initialize r ← r_0
Initialize U, V (e.g., a spectral algorithm)
while not Converged do
    Optimize (3) over U, V until first-order critical
    if (U, V) also second-order critical then
        Converged ← true
    else
        r ← r + 1
        Set (u_{r+1}, v_{r+1}) to be a descent direction
    endif
endw
```

How does this work in practice?

- **First-order** criticality is easily verified and easily reached via a local algorithm (e.g., proximal gradient descent or alternating minimization).
- It is likely computationally **intractable** to verify **second-order** criticality or find a descent direction. For a practical algorithm, we must use a **heuristic**. In [1], we search for a descent direction over one-sparse vectors.

Empirical results

In simulation, Algorithm 1 with alternating minimization and the one-sparse heuristic achieves sample-complexity performance in line with Theorem 1 and comparable to other SOTA practical algorithms (see [1] for comparisons).

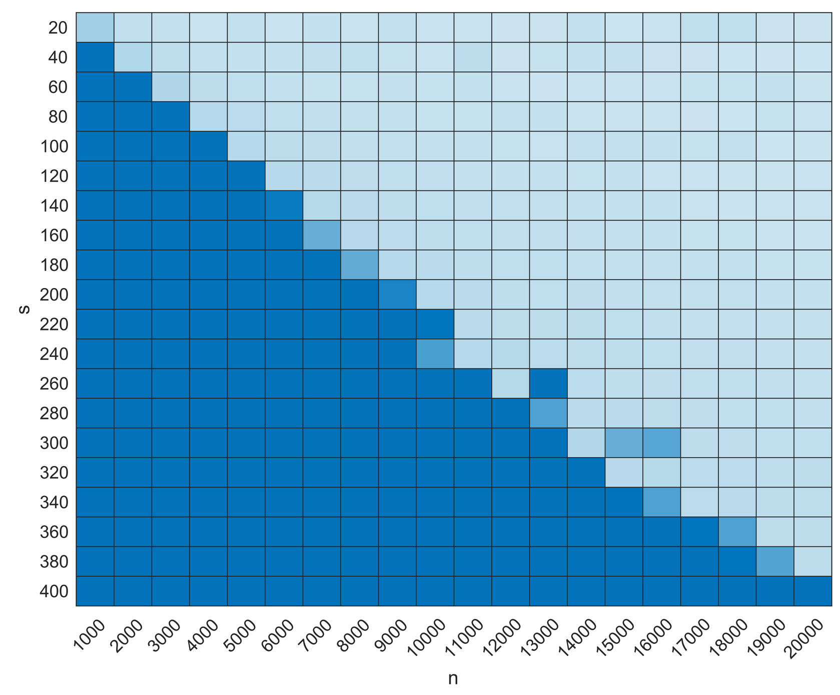

Figure 1: *Error phase diagram (dark = high error, light = low error) of sample size n vs. sparsity s. The required n is (approximately) linear in s, agreeing with Theorem 1.*

Open questions/directions

- **Why** does our heuristic algorithm work well in practice for sparse phase retrieval? Can we prove a theoretical guarantee (thus proving that sparse phase retrieval has **no statistical-computational gap**)?
- How does **problem structure** affect practical solvability (e.g., in comparison with sparse PCA)? **What other problems** are amenable to such practical algorithms?
- What is the **nonsmooth nonconvex landscape** of (2) and (3)? Is it somehow "benign" for certain problems? What theoretical implications would this have for practical **local algorithms**?
- How do we best deal with **symmetry** (e.g., in phase retrieval)? What are the statistical and algorithmic (dis)advantages of enforcing it?

Acknowledgements

The work in [1] was supported in part by the National Science Foundation under Grant CCF-1718771 and Grant CCF-2107455.

References

[1] A. D. McRae, J. Romberg, and M. A. Davenport, "Optimal convex lifted sparse phase retrieval and PCA with an atomic matrix norm regularizer," *IEEE Trans. Inf. Theory*, vol. 69, no. 3, pp. 1866–1882, 2023.

[2] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2886–2908, 2015.

[3] M. Kliesch, S. J. Szarek, and P. Jung, "Simultaneous structures in convex signal recovery—revisiting the convex combination of norms," *Front. Appl. Math. Stat.*, vol. 5, 2019.

[4] T. Wang, Q. Berthet, and R. J. Samworth, "Statistical and computational trade-offs in estimation of sparse principal components," *Ann. Stat.*, vol. 44, no. 5, pp. 1896–1930, 2016.

[5] B. D. Haeffele and R. Vidal, "Structured low-rank matrix factorization: Global optimality, algorithms, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 6, pp. 1468–1482, 2020.

Download

MATERIAL POSTER