First-order methods for low-rank matrix factorization applied to informed source separation

Augustin Lefèvre[1] François Glineur[1,2]
{augustin.lefevre,francois.glineur}@uclouvain.be

[1] Université catholique de Louvain - ICTEAM Institute
Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve - Belgium
[2] Université catholique de Louvain - CORE
Voie du Roman Pays 34, B-1348 Louvain-la-Neuve - Belgium

Abstract: We study a convex formulation of low-rank matrix factorization in the case where additional information on the factors is known. Our formulation is typically adapted to source separation scenarii, where additional information on the sources may be provided by an expert. Our formulation promotes low-rank with a nuclear-norm based penalty. As it is non-smooth, generic first-order algorithms exhibit slow convergence. We introduce and compare two algorithms that fully exploit problem structure while keeping a low computational cost per iteration: minimum-norm subgradient descent and accelerated gradient with a smoothing.

Keywords: source separation, inverse problem, machine learning, subgradient, smoothing.

1 Low-rank matrix factorization and informed source separation

Given a matrix of observations $Y \in \mathbb{R}^{F\times N}$, we assume that Y is a sum of G low-rank contributions perturbed by some noise, i.e. $Y \approx \sum_{g=1}^{G} X_g$ where $X_g \in \mathbb{R}_+^{F\times N}$ is the contribution of source g. The informed source separation problem consists in identifying contributions X_g with the additional knowledge that some entries in some of the contributions are equal to zero.

Matrix factorization is an essential building block in source separation methods. Indeed, as we seek to represent Y as a sum of low-rank matrices, an equivalent problem is to express Y approximately as a low-rank product of factors $D \in \mathbb{R}^{F\times K}$ and corresponding activation coefficients $A \in \mathbb{R}^{K\times N}$ (where the inner size K is the sum of the ranks of the contributions). More specifically, one seeks to minimize a suitable norm of the difference $\|Y - DA\|$. Unfortunately, the problem of identifying the best factors D and A is nonconvex and multimodal, so that in practice only local solutions can be expected to be found in reasonable time (i.e. estimates of D and A in the neighbourhood of which no improvement can be made).

2 A nuclear norm-based nonsmooth convex reformulation

In this work, we substitute the above nonconvex problem with a convex reformulation. A well-known technique to obtain low-rank estimates for the sources X_g is to penalize $\|X_g\|_*$, the nuclear norm of X_g, i.e. the sum of its singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_F$. This leads us to consider the following problem, introduced earlier in [1] :

$$\begin{aligned} \min_{X} \quad & \tfrac{1}{2}\|Y - \textstyle\sum_{g=1}^{G} X_g\|_F^2 + \lambda \sum_{g=1}^{G} \|X_g\|_* \\ \text{subject to} \quad & X_g \geq 0 \text{ and } M_g \cdot X_g = 0, \end{aligned} \tag{1}$$

where $A \cdot B$ denotes the coefficientwise product of A and B, and matrices M_g enforce the constraint that some entries of X_g are equal to zero ($M_{g,fn} \neq 0$ for some source g and coordinates (f, n) implies $X_{g,fn} = 0$). Problem 1 is convex, because we have replaced the requirement that each source term X_g be low-rank by a penalty term $\psi(X) = \lambda \sum_g \|X_g\|_*$ favoring low-rank solutions. Convexity is desirable because it means that global solutions may be reached from any initial point, without recourse to extensive sampling of the search space $\mathbb{R}^{F\times N\times G}$.

3 Algorithms for nuclear norm minimization

Because of the nuclear norm penalty term, the objective function in Problem 1 is non-smooth. In previous experiments with a projected subgradient algorithm [1], we observed that satisfactory source separation could be obtained provided the step size is carefully selected, at the price of a relatively slow convergence. The goal of this work is to introduce and compare algorithms that enjoy faster convergence rates than the projected subgradient. Such an acceleration is made possible by

exploiting the specific structure of the nuclear norm.

A general non-smooth minimization method relies on an arbitrary choice of a subgradient at each iteration, which might not provide a descent direction. In our situation, the subdifferential of the objective function may be described completely, so we can make our choice more wisely, as described in Section 3.2. We also propose in Section 3.3 a second approach that applies an optimal accelerated gradient method to a smooth approximation of the nuclear norm [3].

3.1 Existing work

Besides the projected subgradient method mentioned above, several algorithms for the minimization of the nuclear norm subject to linear equality constraints have been proposed in the literature. For small problems, the authors of [4] show that nuclear norm minimization problems may be reformulated as semidefinite programs (SDP), for which interior-point algorithms with superlinear convergence are available. However, for the purpose of source separation, interior-point methods are too expensive as they require storage and inversion of matrices of size $(F \times N)^2$, with typical dimensions $F \simeq 500$, $N \simeq 10^3$. Another category of algorithms proposed in [2] is based on applying an augmented Lagrangian technique to explicit individual factorizations of the source terms $X_g = D_g A_g$.

3.2 Subgradient methods

Let $\partial f(X)$ be the subdifferential of f at X. Since the objective function f is convex, it also admits directional derivatives $f'(X; D)$ in every direction, and we have $f'(X; D) = \max\{\langle U, D\rangle, U \in \partial f(V)\}$ (we use the standard inner product, that can be defined as $\langle U, V\rangle = \sum_{g=1}^{G} \operatorname{Tr} U_g^\top V_g$).

General subgradient methods for minimizing f successively pick a subgradient $G \in \partial f(V)$ at a given point X, moving X along the direction $-G$ with an appropriate choice of the step size, and projecting X on the set of constraints, if any. However, while the (projected) gradient is always a descent direction when f is differentiable, it is no longer the case for an arbitrary choice of a subgradient. Fortunately, this situation can be improved by considering instead the (opposite of the) minimum norm subgradient $\arg\min\{\|Z\|, Z \in \partial f(X)\}$. In the unconstrained case, this provides the steepest descent direction, which means we can ensure decrease of the objective function with a suitable choice of step size. In our case, additional care must be taken to select a *feasible* descent direction. As computing the minimum norm subgradient implies roughly twice the computational cost of an arbitrary subgradient, it seems worthwhile to compare its merit experimentally.

3.3 Smoothing-based gradient methods

Nesterov [3] showed it is possible to obtain a faster convergence rate for some particular classes of nonsmooth functions (which our problem fits), by applying an accelerated gradient method to a smooth approximation f_μ of the objective function. In our case, we replace the nuclear norm term by :

$$\|X\|_{*,\mu} = \sum_{f=1}^{F} h_\mu(\sigma_f) \qquad h_\mu(x) \triangleq \begin{cases} \frac{x^2}{2\mu} & \text{if } x \leq \mu \\ x - \frac{\mu}{2} & \text{if } x > \mu \end{cases}$$

It can be shown than when $\mu > 0$, the smoothed objective function f_μ has Lipschitz continuous gradient with constant $L_\mu = G + \frac{\lambda}{\mu}$. Following [3] we minimize it with an accelerated gradient method with fixed step size. Moreover, we have $f_\mu(X) \leq f(X) \leq f_\mu(X) + \frac{\mu}{2}$, so that parameter μ trades off the magnitude of the Lipschitz constant L_μ (and hence the rate of convergence of the fast gradient algorithm) with the quality of the approximation (hence its error). Values and first-order derivatives of $\|X\|_{*,\mu}$ are obtained by computing the SVD of X. Hence we can implement a fast gradient method with the same iteration cost as the projected subgradient descent, but with a superior theoretical convergence rate.

4 Experimental results

We compare the subgradient and the smoothing approaches described in Sections 3.2 and 3.3 in an informed source separation experiment involving four musical pieces (14 seconds each). We first show the superiority of the minimum norm subgradient over an arbitrary subgradient, even after taking its higher cost into account. We then observe that, for appropriate values of μ, the accelerated gradient with smoothing achieves both faster decrease of the objective function and better quality solutions than competing approaches.

Acknowledgments

This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

References

[1] A. Lefèvre, F. Glineur, and P.-A. Absil. A nuclear norm-based convex formulation for informed source separation. Technical Report 1212.3119, arXiv, 2012.

[2] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 2009.

[3] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 2005.

[4] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review.*, 2010.