

PROVABLY CONVERGENT NONCONVEX ALGORITHM FOR VOLUME OPTIMIZATION-BASED LATENT COMPONENT ANALYSES

Anonymous authors

Paper under double-blind review

ABSTRACT

We present an algorithm that aims at solving a family of nonconvex optimization problems with convergence guarantee to a global optimum. This family of nonconvex optimization problems share the formulation of maximizing the volume of a matrix subject to linear constraints. Problems of this form have found a lot of applications in unsupervised learning and representation learning, especially when identifiability of the latent representation is important for the task. Specific examples based on the types of constraints include bounded component analysis, sparse component analysis (complete dictionary learning), nonnegative component analysis (nonnegative matrix factorization), and admixture component analysis, to name a few. Computationally, the problem is hard because of the nonconvex objective. An algorithm based on linearized ADMM is proposed for these problems. Although a similar algorithm has appeared in the literature, we note that a small modification has to be made in order to guarantee that the algorithm provably converges even for non-convex problems. Our main contribution is convergence guarantee to a global optimum at a sublinear rate. We do assume some mild conditions on the initialization, but our numerical experiments indicate that these initialization conditions are very easy to satisfy.

1 INTRODUCTION

Many problems in unsupervised representation learning require identifying the latent components from data with structural assumptions/constraints on the latent representation. While some works show that volume optimization can guarantee identifiability of the latent components under mild conditions, the provably convergent algorithm is still missing for any of these problems. To bridge this gap, we propose an algorithm that aims at solving the following (family of) nonconvex optimization problem:

$$\underset{\mathbf{W}}{\text{minimize}} \quad -\frac{1}{2} \log \det \mathbf{W}\mathbf{W}^\top + g(\mathbf{W}\mathbf{X}), \quad (1)$$

where g is some regularization function to enforce or promote certain problem-specific structure to the transformation $\mathbf{W}\mathbf{X}$ to a lower dimension, therefore $\mathbf{W}$ should be a wide matrix with full row rank. This type of problems naturally arises in many unsupervised learning problems, where data samples are stacked in the columns of $\mathbf{X}$ and the goal is to find a linear transformation $\mathbf{W}$ that makes the learned representation/embedding $\mathbf{W}\mathbf{X}$ satisfy some structural requirements. A classical example is independent component analysis (ICA), where ideally $g(\mathbf{W}\mathbf{X})$ should be a function to enforce that the rows of $\mathbf{W}\mathbf{X}$ are statistically independent (Comon, 1994; Hyvärinen et al., 2004), although in practice it is fairly hard to find such a well-defined function, especially if working with a finite set of data $\mathbf{X}$. Ever since, some structural constraints/regularizations that are more suitable for finite data sets with explicitly well-defined functional forms have been studied, including:

- Bounded component analysis (BCA) (Tatli & Erdogan, 2021; Hu & Huang, 2023b). In this case g is an indicator function defined as

$$g(\mathbf{S}) = \begin{cases} 0 & |S_{ij}| \leq 1 \text{ for all } i, j \\ +\infty & \text{otherwise.} \end{cases}$$

Notice that $|\cdot|$ could mean magnitude of a complex number if data are complex, in which case replace matrix transpose $^\top$ with Hermitian transpose $^\mathsf{H}$ (Hu & Huang, 2024).

- Sparse component analysis (SCA) (Hu & Huang, 2023a), also known as (complete) dictionary learning. Here g is simply the (vector) ℓ_1 norm, i.e., sum of absolute values

$$g(\mathbf{S}) = \|\mathbf{S}\|_1 = \sum_{i,j} |S_{ij}|.$$

Again, it can be extended to complex numbers with ease (Sun & Huang, 2024). We will not consider overcomplete dictionaries in this paper, although a related formulation has appeared (Sun & Huang, 2025).

- Nonnegative component analysis (NCA) (Fu et al., 2018), which is a variant of nonnegative matrix factorization (NMF) with only one of the factors to be nonnegative:

$$g(\mathbf{S}) = \begin{cases} \sum_{i,j} S_{ij} & S_{ij} \geq 0 \text{ for all } i, j \\ +\infty & \text{otherwise.} \end{cases}$$

- Admixture component analysis (ACA). In this case, each embedding $\mathbf{W}\mathbf{x}_i$ represent the abundance/admixture of the components, meaning the coefficients are nonnegative and sum to one, so g is again an indicator function

$$g(\mathbf{S}) = \begin{cases} 0 & \mathbf{S} \geq 0, \mathbf{I}^\top \mathbf{S} = \mathbf{I}^\top \\ +\infty & \text{otherwise.} \end{cases}$$

ACA has been widely adopted in topic models (Blei, 2012) and hyperspectral unmixing (Ma et al., 2013).

- Principal component analysis (PCA). We throw in PCA here for completeness, even though it is unnecessary as there are countless algorithms that efficiently solves it. We show in the supplementary that if $g(\mathbf{S}) = \|\mathbf{S}\|_{\mathbb{F}}^2$, the solution of (1) recovers that of PCA, exhibiting versatility of the framework.

1.1 IDENTIFIABILITY VIA VOLUME OPTIMIZATION

For a wide matrix $\mathbf{W}$ with linearly independent rows, $\sqrt{\det \mathbf{W}\mathbf{W}^\top}$ is sometimes called the volume of the matrix (Ben-Israel, 1992) with applications in change-of-variable integration (Ben-Israel, 1999) and probability (Ben-Israel, 2000). In particular, if $\mathbf{s} \in \mathbb{R}^k$ is a latent random variable with distribution $p(\mathbf{s})$, and we observe $\mathbf{x} = \mathbf{A}\mathbf{s}$, then the distribution of $\mathbf{x}$ admits the following relation

$$p(\mathbf{x}) = (\det \mathbf{A}^\top \mathbf{A})^{-1/2} p(\mathbf{s}).$$

Therefore, applying a change-of-variable $\mathbf{A} = \mathbf{W}^\top$, formulation (1) is exactly the maximum likelihood estimate with the following latent distributions:

- PCA: standard Gaussian distribution
- SCA: standard Laplacian distribution
- NCA: standard exponential distribution
- BCA: uniform distribution in the hypercube
- ACA: uniform Dirichlet distribution

Even for general ICA models, as long as the type of latent variables are known, one can write down the exact maximum likelihood formulation in the form of (1), for instance the Cauchy distribution with $g(s) = -\log(1 + s^2)$.

The bigger benefit is identifiability, with the exception of PCA. Suppose the observations are indeed generated from the model $\mathbf{X} = \mathbf{A}\mathbf{S}$ with unknown $\mathbf{A}$ and $\mathbf{S}$, in many cases one is interested in exactly recover $\mathbf{S}$ up to trivial ambiguities such as row permutation and scaling. Historically, the original motivation to propose ICA was exactly because PCA is not identifiable. However, exact identifiability of ICA requires statistical independence, which is not clear how to quantify with finite data. The volume optimization framework for structured latent component analysis has been shown to guarantee identifiability under realistic assumptions. In particular, denote the true latent components as $\mathbf{S}_\natural$ and a solution to (1) as $\mathbf{S}_\star = \mathbf{W}_\star \mathbf{X}$, then $\mathbf{S}_\star$ is a row permutation and scaling of $\mathbf{S}_\natural$ if

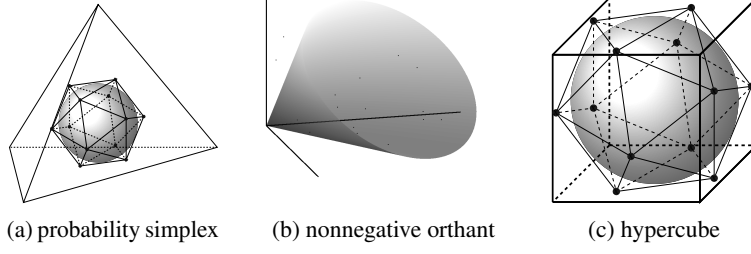


Figure 1: Illustrations of the sufficiently scattered condition in the (a) probability simplex, (b) nonnegative orthant, and (c) hypercube in $\mathbb{R}^3$.

- ACA: convex hull of $\mathbf{S}_h$, denoted as $\text{conv}(\mathbf{S}_h) = \{\mathbf{S}_h \boldsymbol{\theta} \mid \boldsymbol{\theta} \geq 0, \mathbf{I}^\top \boldsymbol{\theta} = 1\}$, is sufficiently scattered in the probability simplex (Fu et al., 2015; Lin et al., 2015)

$$\{\mathbf{x} \in \mathbb{R}^k \mid \|\mathbf{x}\|_2 \leq 1/\sqrt{k-1}, \mathbf{I}^\top \mathbf{x} = 1\} \subset \text{conv}(\mathbf{S}_h).$$

- NCA: conic hull of $\mathbf{S}_h$, denoted as $\text{cone}(\mathbf{S}_h) = \{\mathbf{S}_h \boldsymbol{\theta} \mid \boldsymbol{\theta} \geq 0\}$, is sufficiently scattered in the nonnegative orthant (Huang et al., 2013; Fu et al., 2018)

$$\{\mathbf{x} \in \mathbb{R}^k \mid \|\mathbf{x}\|_2 \leq \mathbf{I}^\top \mathbf{x} / \sqrt{k-1}\} \subset \text{cone}(\mathbf{S}_h).$$

- BCA: disked hull of $\mathbf{S}_h$, denoted as $\text{disk}(\mathbf{S}_h) = \{\mathbf{S}_h \boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_1 \leq 1\}$, is sufficiently scattered in the hypercube (Tatli & Erdogan, 2021; Hu & Huang, 2023b)

$$\{\mathbf{x} \in \mathbb{R}^k \mid \|\mathbf{x}\|_2 \leq 1\} \subset \text{disk}(\mathbf{S}_h).$$

- SCA: cellular hull of $\tilde{\mathbf{S}}_h$, where $\tilde{\mathbf{S}}_h$ is obtained from rescaling the rows of $\mathbf{S}_h$ to unit ℓ_1 norm, denoted as $\text{cell}(\tilde{\mathbf{S}}_h) = \{\tilde{\mathbf{S}}_h \boldsymbol{\theta} \mid \|\boldsymbol{\theta}\|_\infty \leq 1\}$, is sufficiently scattered in the hypercube (Hu & Huang, 2023a)

$$\{\mathbf{x} \in \mathbb{R}^k \mid \|\mathbf{x}\|_2 \leq 1\} \subset \text{cell}(\tilde{\mathbf{S}}_h).$$

Notice that the complex case for BCA (Hu & Huang, 2024) and SCA (Sun & Huang, 2024) are also true.

Illustrations of these various sufficiently scattered conditions are shown in Figure 1. More details on identifiability of these models can be found in the aforementioned references and therein.

1.2 THIS PAPER

This work focuses on algorithmic analysis for (1). First we note that (1) is nonconvex due to the volume term $-\log \det \mathbf{W}\mathbf{W}^\top$; although log-determinant of a positive definite matrix is concave, it is not the case when the positive definite matrix is parameterized by $\mathbf{W}\mathbf{W}^\top$. On the other hand, g is usually a convex function or indicator function of a convex set, and usually admits an efficient proximal/projection operator. Some prior work on algorithmic design for (1) are reviewed in the supplementary. In this paper we propose a new algorithm based on the linearized alternating direction method of multipliers (L-ADMM) with a constant step size γ as follows

$$\begin{cases} \mathbf{W}_{t+1} \leftarrow (\mathbf{S}_t + \mathbf{U}_t + \gamma \mathbf{S}_t^{\dagger\top}) \mathbf{X}^\dagger \\ \mathbf{S}_{t+1} \leftarrow \text{Prox}_{\gamma g}(\mathbf{W}_{t+1} \mathbf{X} - \mathbf{U}_t) \\ \mathbf{U}_{t+1} \leftarrow \mathbf{U}_t + \mathbf{S}_{t+1} - \mathbf{W}_{t+1} \mathbf{X} \end{cases} \quad (2)$$

Derivation of the algorithm will be explained in detail in the sequel. An immediate observation is that (2) is very easy to implement in practice, as it only involves basic operations such as matrix multiplication, pseudo-inverse, and proximal operators—for the aforementioned specific examples, their corresponding proximal operators are as simple as truncations (for BCA) or soft-thresholding (for SCA). Despite its simplicity, its performance is surprisingly good, as will be demonstrated in the experiment section. Recognizing that we are trying to solve a nonconvex problem (1), with some

cases shown to be NP-hard (Packer, 2002), those numerical results seem incredible as it almost always achieves global optimality within a short amount of time. This inspires us to further investigate the algorithmic behavior of (2) to explain its excellent performance. In this paper we make the following contributions:

1. We proposed a general volume-optimization framework that includes the aforementioned problems (e.g., BCA/SCA/NCA/ACA, which are generally non-convex and NP-hard). More importantly, our newly designed algorithm is the first method with provable convergence to global optimum.
2. Compared to state-of-the-art baselines (BCD, FW, ADMM algorithms) that tailor designed for each identifiable problems, our new method is either significantly faster or at least comparable. This is also a non-trivial contribution on top of our theoretical results.

We want to remark our major technical significances. We introduce one version of linearized alternating direction method of multipliers (L-ADMM) that has not been proposed to the best of our knowledge. The reason we choose this version is thanks to its simplicity in proving convergence for convex problems, with minimal assumptions. It inspired us to generalize into non-convex problem.

Recognizing that convexity of the objective function is only used to form first-order lowerbounds and the other point is always a global optimum, a similar inequality is established for (2) even though $-\log \det \mathbf{W}\mathbf{W}^\top$ is not convex. The claim is that if $\mathbf{X}$ is indeed generated from an identifiable model $\mathbf{X} = \mathbf{A}_\sharp \mathbf{S}_\sharp$ and the initialization $\mathbf{S}_0$ satisfies

$$\log |\det \mathbf{S}_0 \mathbf{S}_\sharp^\dagger| \leq \max_{\Pi, \mathbf{D}} \text{Tr } \Pi \mathbf{D} \mathbf{S}_0 \mathbf{S}_\sharp^\dagger - k, \quad (3)$$

where Π are permutation matrices and $\mathbf{D}$ are diagonal matrices with ± 1 on the diagonal, then L-ADMM converges to a global optimum. Although we do not provide an initialization scheme that always satisfies (3), our numerical experiments suggest that it is very easy to satisfy (3) with random initializations, thus explaining the above mentioned excellent performance.

2 LINEARIZED ADMM

Alternating direction method of multipliers (ADMM) has been a widely adopted algorithm in machine learning (Boyd et al., 2011). Consider a generic linearly constrained optimization problem

$$\begin{aligned} & \underset{\mathbf{w}, \mathbf{s}}{\text{minimize}} \quad f(\mathbf{w}) + g(\mathbf{s}) \\ & \text{subject to} \quad \mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{s} = \mathbf{c}, \end{aligned} \quad (4)$$

ADMM takes the following iterative form

$$\begin{cases} \mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} f(\mathbf{w}) + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{s}_t - \mathbf{c} + \mathbf{u}_t\|^2 \\ \mathbf{s}_{t+1} \leftarrow \arg \min_{\mathbf{s}} g(\mathbf{s}) + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s} - \mathbf{c} + \mathbf{u}_t\|^2 \\ \mathbf{u}_{t+1} \leftarrow \mathbf{u}_t + \mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} \end{cases}$$

where γ is the step size. One of the possible downside is that it involves potentially nontrivial minimizations within each iteration that may not admit to closed-form updates. Suppose this is true for the update of $\mathbf{w}$, then one way to mitigate this is to linearize $f(\mathbf{w})$ in each iteration, leading to Linearized ADMM (L-ADMM)

$$\begin{cases} \mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \nabla f(\mathbf{A}^\dagger(\mathbf{c} - \mathbf{B}\mathbf{s}_t))^\top \mathbf{w} + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{s}_t - \mathbf{c} + \mathbf{u}_t\|^2 \\ \mathbf{s}_{t+1} \leftarrow \arg \min_{\mathbf{s}} g(\mathbf{s}) + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s} - \mathbf{c} + \mathbf{u}_t\|^2 \\ \mathbf{u}_{t+1} \leftarrow \mathbf{u}_t + \mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} \end{cases} \quad (5)$$

Note that the first order Taylor approximation of f is taken at $\mathbf{A}^\dagger(\mathbf{c} - \mathbf{B}\mathbf{s}_t)$, not $\mathbf{w}_t$ as in (Hu & Huang, 2023a). To the best of our knowledge, this type of linearization has not appeared in the literature, among the various inexact ADMM variants (He et al., 2002; Ng et al., 2011; Lin et al., 2017; Gao et al., 2018; Lu et al., 2021). However, we find that this version leads to a relatively simple convergence proof for convex problems with minimal assumptions, i.e., f is Lipschitz smooth and

columns of $\mathbf{A}$ are linearly independent. This would be a particularly useful property for us to further apply it to nonconvex problems that are of interest in this work, as we will see shortly.

Here we provide a concise convergence proof for L-ADMM assuming both f and g are convex, while f is also Lipschitz smooth. Our analysis differs from most existing work by employing a different Lyapunov function, which will be shown to converge to zero. Consider the Lagrangian function of (4)

$$L(\mathbf{w}, \mathbf{s}, \boldsymbol{\lambda}) = f(\mathbf{w}) + g(\mathbf{s}) + \boldsymbol{\lambda}^\top (\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{s} - \mathbf{c}).$$

If strong duality holds, then we have

$$L(\mathbf{w}, \mathbf{s}, \boldsymbol{\lambda}_\star) \geq L(\mathbf{w}_\star, \mathbf{s}_\star, \boldsymbol{\lambda}_\star) = f(\mathbf{w}_\star) + g(\mathbf{s}_\star),$$

where $(\mathbf{w}_\star, \mathbf{s}_\star)$ and $\boldsymbol{\lambda}_\star$ are optimal primal and dual variables, since $(\mathbf{w}_\star, \mathbf{s}_\star)$ minimizes $L(\mathbf{w}, \mathbf{s}, \boldsymbol{\lambda}_\star)$ and they are also feasible, i.e.,

$$\mathbf{A}\mathbf{w}_\star + \mathbf{B}\mathbf{s}_\star = \mathbf{c}. \quad (6)$$

We will show in the sequel that $L(\mathbf{w}_t, \mathbf{s}_t, \boldsymbol{\lambda}_\star) - L(\mathbf{w}_\star, \mathbf{s}_\star, \boldsymbol{\lambda}_\star)$, which is a nonnegative sequence, goes to zero as $t \rightarrow \infty$.

Theorem 2.1. *Suppose f and g are convex functions, f is Lipschitz smooth, meaning there exists a constant M such that for all $\mathbf{w}$ and $\tilde{\mathbf{w}}$ in the domain of f*

$$\|\nabla f(\mathbf{w}) - \nabla f(\tilde{\mathbf{w}})\| \leq M\|\mathbf{w} - \tilde{\mathbf{w}}\|,$$

and matrix $\mathbf{A}$ has full column rank: there exists a constant μ such that

$$\mathbf{A}^\top \mathbf{A} \succeq \mu \mathbf{I}.$$

Then with $\gamma \leq \mu/M$, we have

$$\min_{t=1, \dots, T} L(\mathbf{w}_t, \mathbf{s}_t, \boldsymbol{\lambda}_\star) - L(\mathbf{w}_\star, \mathbf{s}_\star, \boldsymbol{\lambda}_\star) \leq \frac{1}{T} \left(\frac{1}{2\gamma} \|\mathbf{A}(\mathbf{w}_0 - \mathbf{w}_\star) + \mathbf{u}_0 - \gamma \boldsymbol{\lambda}_\star\|^2 \right). \quad (7)$$

Proof. First we rearrange the order of L-ADMM to

$$\begin{cases} \mathbf{s}_{t+1} \leftarrow \arg \min_{\mathbf{s}} g(\mathbf{s}) + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_t + \mathbf{B}\mathbf{s} - \mathbf{c} + \mathbf{u}_t\|^2 \\ \mathbf{u}_{t+1} \leftarrow \mathbf{u}_t + \mathbf{A}\mathbf{w}_t + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} \\ \mathbf{w}_{t+1} \leftarrow \arg \min_{\mathbf{w}} \nabla f(\mathbf{A}^\dagger(\mathbf{c} - \mathbf{B}\mathbf{s}_{t+1}))^\top \mathbf{w} + \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w} + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} + \mathbf{u}_{t+1}\|^2 \end{cases} \quad (8)$$

This is equivalent to decreasing the iteration index of $\mathbf{w}$ by 1.

The update of $\mathbf{s}$ implies

$$-\frac{1}{\gamma} \mathbf{B}^\top (\mathbf{A}\mathbf{w}_t + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} + \mathbf{u}_t) \in \partial g(\mathbf{s}_{t+1}),$$

for convex g we have

$$g(\mathbf{s}_{t+1}) - g(\mathbf{s}_\star) \leq \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_t + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} + \mathbf{u}_t)^\top \mathbf{B}(\mathbf{s}_\star - \mathbf{s}_{t+1}). \quad (9)$$

The update of $\mathbf{u}_{t+1}$ in (5) implies

$$\mathbf{B}\mathbf{s}_{t+1} = \mathbf{c} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t.$$

Substituting it into (9) together with $\mathbf{B}\mathbf{s}_\star = \mathbf{c} - \mathbf{A}\mathbf{w}_\star$ from (6) gives

$$g(\mathbf{s}_{t+1}) - g(\mathbf{s}_\star) \leq \frac{1}{\gamma} \mathbf{u}_{t+1}^\top (\mathbf{A}\mathbf{w}_t - \mathbf{A}\mathbf{w}_\star + \mathbf{u}_t - \mathbf{u}_{t+1}). \quad (10)$$

Denote $\hat{\mathbf{w}}_t = \mathbf{A}^\dagger(\mathbf{c} - \mathbf{B}\mathbf{s}_{t+1})$. From convexity and Lipschitz-smoothness of f , we have

$$\begin{aligned} f(\mathbf{w}_{t+1}) &\leq f(\hat{\mathbf{w}}_t) + \nabla f(\hat{\mathbf{w}}_t)^\top (\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2, \\ f(\mathbf{w}_\star) &\geq f(\hat{\mathbf{w}}_t) + \nabla f(\hat{\mathbf{w}}_t)^\top (\mathbf{w}_\star - \hat{\mathbf{w}}_t). \end{aligned} \quad (11)$$

From the update of $\mathbf{w}$ in (5), we have

$$\nabla f(\hat{\mathbf{w}}_t) = -\frac{1}{\gamma} \mathbf{A}^\top (\mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c} + \mathbf{u}_{t+1}) = -\frac{1}{\gamma} \mathbf{A}^\top (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + 2\mathbf{u}_{t+1} - \mathbf{u}_t)$$

All three equations above shows

$$f(\mathbf{w}_{t+1}) - f(\mathbf{w}_\star) \leq \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + 2\mathbf{u}_{t+1} - \mathbf{u}_t)^\top \mathbf{A}(\mathbf{w}_\star - \mathbf{w}_{t+1}) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2. \quad (12)$$

Combining (10) and (12) gives

$$\begin{aligned} & f(\mathbf{w}_{t+1}) + g(\mathbf{s}_{t+1}) - f(\mathbf{w}_\star) - g(\mathbf{s}_\star) \\ & \leq \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + 2\mathbf{u}_{t+1} - \mathbf{u}_t)^\top \mathbf{A}(\mathbf{w}_\star - \mathbf{w}_{t+1}) + \frac{1}{\gamma} \mathbf{u}_{t+1}^\top (\mathbf{A}\mathbf{w}_t - \mathbf{A}\mathbf{w}_\star + \mathbf{u}_t - \mathbf{u}_{t+1}) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \\ & = \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t)^\top \mathbf{A}(\mathbf{w}_\star - \mathbf{w}_{t+1}) + \frac{1}{\gamma} \mathbf{u}_{t+1}^\top (\mathbf{A}\mathbf{w}_t - \mathbf{A}\mathbf{w}_{t+1} + \mathbf{u}_t - \mathbf{u}_{t+1}) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \\ & = \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t)^\top (\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_{t+1} - \mathbf{u}_{t+1}) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \end{aligned}$$

Adding both sides by

$$\boldsymbol{\lambda}_\star^\top (\mathbf{A}\mathbf{w}_{t+1} + \mathbf{B}\mathbf{s}_{t+1} - \mathbf{c}) = \boldsymbol{\lambda}_\star^\top (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t),$$

where the right-hand-side is obtained from the update rule of $\mathbf{u}_{t+1}$, we get

$$\begin{aligned} & L(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, \boldsymbol{\lambda}_\star) - L(\mathbf{w}_\star, \mathbf{s}_\star, \boldsymbol{\lambda}_\star) \\ & \leq \frac{1}{\gamma} (\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t)^\top (\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_{t+1} + \gamma \boldsymbol{\lambda}_\star - \mathbf{u}_{t+1}) + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \\ & = \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_t + \gamma \boldsymbol{\lambda}_\star - \mathbf{u}_t\|^2 - \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_{t+1} + \gamma \boldsymbol{\lambda}_\star - \mathbf{u}_{t+1}\|^2 \\ & \quad - \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t\|^2 + \frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \end{aligned} \quad (13)$$

Notice that we defined $\hat{\mathbf{w}}_t = \mathbf{A}^\dagger (\mathbf{c} - \mathbf{B}\mathbf{s}_{t+1}) = \mathbf{A}^\dagger (\mathbf{A}\mathbf{w}_t + \mathbf{u}_t - \mathbf{u}_{t+1})$, and if $\mathbf{A}^\top \mathbf{A} \succeq \mu \mathbf{I}$ and $\gamma \leq \mu/M$, then

$$\frac{M}{2} \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2 \leq \frac{M}{2\mu} \|\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t\|^2 \leq \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_{t+1} - \mathbf{A}\mathbf{w}_t + \mathbf{u}_{t+1} - \mathbf{u}_t\|^2.$$

This means the last line of (13) is nonpositive, and thus

$$\begin{aligned} & L(\mathbf{w}_{t+1}, \mathbf{s}_{t+1}, \boldsymbol{\lambda}_\star) - L(\mathbf{w}_\star, \mathbf{s}_\star, \boldsymbol{\lambda}_\star) \\ & = \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_t + \gamma \boldsymbol{\lambda}_\star - \mathbf{u}_t\|^2 - \frac{1}{2\gamma} \|\mathbf{A}\mathbf{w}_\star - \mathbf{A}\mathbf{w}_{t+1} + \gamma \boldsymbol{\lambda}_\star - \mathbf{u}_{t+1}\|^2 \end{aligned} \quad (14)$$

Taking the summation of (14) with $t = 1, \dots, T$, omit the negative terms on the right-hand-side, and replace each $L(\mathbf{w}_t, \mathbf{s}_t, \boldsymbol{\lambda}_\star)$ with their min on the left-hand-side, gives us (7). $\square$

The proof clearly explains why the linearization is taken at $\hat{\mathbf{w}}_t = \mathbf{A}^\dagger (\mathbf{c} - \mathbf{B}\mathbf{s}_{t+1})$: the Lipschitz-smoothness property introduces a nonnegative term $(M/2) \|\mathbf{w}_{t+1} - \hat{\mathbf{w}}_t\|^2$ that needs to be eliminated in order to guarantee convergence, and the last line of (13) provides the choice of $\hat{\mathbf{w}}_t$ to achieve this. Numerically, this result hints that one could precondition the matrix $\mathbf{A}$ for better numerical performance (and a larger range of choice for γ). We find that orthogonalizing the columns of $\mathbf{A}$ works really well in practice; since the update of $\mathbf{w}$ involves solving a least squares problem, a linear transformation of $\mathbf{w}$ is not going to change the complexity of its update. The proof also suggests that the linearization step should be done first (or, equivalently, immediately after the dual update), not the other way around, otherwise convergence may not be guaranteed. We obtained a sublinear convergence rate of $1/T$ without assuming strong convexity, which is similar to proximal gradient descent. If we assume f to be strongly convex, one may expect to achieve a faster linear convergence rate with ease, but that is beyond the scope of this work.

Remark. Throughout the proof, convexity of f and g are only invoked to form the linear lower-bounds (9) and (11); moreover, one of the points is always at the optimum $\mathbf{s}_\star$ or $\mathbf{w}_\star$. This observation hints that even if, say, f is not convex, as long as the iterates of L-ADMM generates $\hat{\mathbf{w}}_t$ that satisfies (11), then global optimality can still be guaranteed. This is going to be the main focus when applying L-ADMM to nonconvex problems.

3 L-ADMM FOR VOLUME OPTIMIZATION-BASED LATENT COMPONENT ANALYSES

Before applying L-ADMM, we first reformulate (1) by introducing auxiliary variable $\mathbf{S} = \mathbf{W}\mathbf{X}$ as

$$\begin{aligned} \underset{\mathbf{W}, \mathbf{S}}{\text{minimize}} \quad & -\frac{1}{2} \log \det \mathbf{W}\mathbf{W}^\top + g(\mathbf{S}) \\ \text{subject to} \quad & \mathbf{S} = \mathbf{W}\mathbf{X}. \end{aligned} \quad (15)$$

This formulation has two separate terms for $\mathbf{W}$ and $\mathbf{S}$ coupled by a linear constraint $\mathbf{S} = \mathbf{W}\mathbf{X}$, which is a form that L-ADMM can easily be applied. Its exact form has been presented in (2).

3.1 IMPLEMENTATION DETAILS

The updates of $\mathbf{S}$ and $\mathbf{U}$ follows exactly from the definition of L-ADMM in (5). The update of $\mathbf{W}$ requires a bit more explanation. First of all, the gradient of we know the gradient of the log-determinant objective is $-(\mathbf{W}^\top)^\top$ and since it is evaluated at $\mathbf{S}_t \mathbf{X}^\dagger$, we have that

$$\nabla -\log \det(\mathbf{S}_t \mathbf{X}^\dagger)(\mathbf{S}_t \mathbf{X}^\dagger)^\top = -(\mathbf{S}_t \mathbf{X}^\dagger)^{\dagger\top} = -\mathbf{S}_t^{\dagger\top} \mathbf{X}^\top,$$

where the last equality uses the fact that $(\mathbf{A}\mathbf{B})^\dagger = (\mathbf{A}^\dagger \mathbf{A}\mathbf{B})^\dagger (\mathbf{A}\mathbf{B}\mathbf{B}^\dagger)^\dagger$ (Petersen et al., 2008), which further equals to $\mathbf{B}^\dagger \mathbf{A}^\dagger$ if $\mathbf{A}$ has full column rank and $\mathbf{B}$ has full row rank. The update of $\mathbf{W}_{t+1}$ is explicitly defined as

$$\mathbf{W}_{t+1} = \arg \min_{\mathbf{W}} -\text{Tr} \mathbf{W}^\top \mathbf{S}_t^{\dagger\top} \mathbf{X}^\top + \frac{1}{2\gamma} \|\mathbf{S}_t - \mathbf{W}\mathbf{X} + \mathbf{U}_t\|^2.$$

Setting its gradient equals zero gives the $\mathbf{W}$ update in (2).

As for the proximal operators for various g , it has been studied extensively, so we simply list some of them here without derivations:

$$\begin{aligned} \text{BCA: } \text{Prox}_{\gamma g}(\mathbf{S})_{ij} &= \begin{cases} \text{sign}(S_{ij}), & |S_{ij}| > 1, \\ S_{ij}, & \text{otherwise.} \end{cases} \\ \text{SCA: } \text{Prox}_{\gamma g}(\mathbf{S})_{ij} &= \begin{cases} \text{sign}(S_{ij})(|S_{ij}| - \gamma), & |S_{ij}| > \gamma, \\ 0, & \text{otherwise.} \end{cases} \\ \text{NCA: } \text{Prox}_{\gamma g}(\mathbf{S})_{ij} &= [S_{ij} - \gamma]_+ \\ \text{ACA: } \text{Prox}_{\gamma g}(\mathbf{s}_i) &= [\mathbf{s}_i - \mathbf{v}_i \mathbf{I}]_+ \end{aligned}$$

Simplex projection for ACA is taken column-wise of $\mathbf{S}$ and each $\mathbf{v}_i$ is a scalar to satisfy that $\mathbf{I}^\top [\mathbf{s}_i - \mathbf{v}_i \mathbf{I}]_+ = 1$; there exist several methods that keep the complexity linear, such as bisection (Parikh & Boyd, 2014) or divide-and-conquer (Duchi et al., 2008).

3.2 GLOBAL CONVERGENCE

We now present the main contribution of this work, namely to show that Algorithm (2) converges to a global optimum of (1) under some mild conditions. As formulation (1) is mostly used for identifiable latent component analyses such as BCA or SCA, we focus on the case when $\mathbf{X}$ is generated from $\mathbf{X} = \mathbf{A}_\natural \mathbf{S}_\natural$, where $\mathbf{A}_\natural$ and $\mathbf{S}_\natural$ are the true latent factors, and $\mathbf{S}_\natural$ satisfies one of the sufficiently scattered conditions described in §1.1, therefore any solution to (1) must be a row permutation and sign flip (if BCA or SCA) of $\mathbf{S}_\natural$.

Denote $f(\mathbf{W}) = -\log \det \mathbf{W}\mathbf{W}^\top$. Proof of Theorem 2.1 indicates that convexity of f is only invoked when forming a linear lowerbound (11) at $\mathbf{S}_t \mathbf{X}^\dagger = \mathbf{S}_t (\mathbf{A}_\mathfrak{h} \mathbf{S}_\mathfrak{h})^\dagger$ and $\mathbf{W}_\star = \mathbf{A}_\mathfrak{h}^\dagger$, i.e.,

$$-\frac{1}{2} \log \det (\mathbf{A}_\mathfrak{h}^\dagger \mathbf{A}_\mathfrak{h}^{\dagger\top}) \geq -\frac{1}{2} \log \det (\mathbf{S}_t (\mathbf{A}_\mathfrak{h} \mathbf{S}_\mathfrak{h})^\dagger (\mathbf{A}_\mathfrak{h} \mathbf{S}_\mathfrak{h})^{\dagger\top} \mathbf{S}_t^\top) - \text{Tr } \mathbf{X} \mathbf{S}_t^\top (\mathbf{A}_\mathfrak{h}^\dagger - \mathbf{S}_t \mathbf{X}^\dagger).$$

This simplifies to

$$\log |\det \mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger| \leq \text{Tr } \mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger - k. \quad (16)$$

We should be reminded that (16) does not always hold, but if it was true for all $\mathbf{S}_t$, then it suffices to guarantee that L-ADMM converges to $\mathbf{S}_\mathfrak{h}$. In fact, since any row permutation and sign flip (for BCA and SCA) of $\mathbf{S}_\mathfrak{h}$ is also optimal for (15), we can replace $\mathbf{S}_\mathfrak{h}$ in (16) with $\mathbf{\Pi} \mathbf{D} \mathbf{S}_\mathfrak{h}$ where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix with ± 1 on the diagonal (for NCA and ACA we would have $\mathbf{D} = \mathbf{I}$, but we include this term for completeness), then global convergence can be guaranteed.

To gain some insights about what (16) implies, obviously the inequality holds for positive definite matrices. However, $\mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger$ is not positive definite. While the left-hand-side is related to the log-determinant of its Gram matrix, the right-hand-side only depends on its diagonal values. Therefore, intuitively, (16) would hold if $\mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger$ is “diagonally dominant”, to some extent. Indeed, if $\mathbf{S}_t$ is optimal, then $\mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger = \mathbf{I}$, which is the most diagonally dominant matrix. To put it differently, (16) holds when $\mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger$ is relatively close to $\mathbf{I}$; it does not have to be extremely close like many nonconvex algorithmic analysis requires, just close enough that (16) holds suffices. In other words, (16) defines the “basin of attraction” for the nonconvex problem (15).

Theorem 3.1. *Let $\mathbf{X} = \mathbf{A}_\mathfrak{h} \mathbf{S}_\mathfrak{h}$ where $\mathbf{A}_\mathfrak{h}$ and $\mathbf{S}_\mathfrak{h}$ are the true latent factors, and $\mathbf{S}_\mathfrak{h}$ satisfies one of the sufficiently scattered conditions described in §1.1. Suppose the L-ADMM algorithm (2) is initialized with $\mathbf{S}_0$ that satisfies*

$$\log |\det \mathbf{S}_0 \mathbf{S}_\mathfrak{h}^\dagger| \leq \max_{\mathbf{\Pi}, \mathbf{D}} \text{Tr } \mathbf{\Pi} \mathbf{D} \mathbf{S}_0 \mathbf{S}_\mathfrak{h}^\dagger - k,$$

where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{D}$ is a diagonal matrix with ± 1 on the diagonal, then we have

$$\min_{t=1, \dots, T} L(\mathbf{W}_t, \mathbf{S}_t, \boldsymbol{\lambda}_\star) - L(\mathbf{W}_\star, \mathbf{S}_\star, \boldsymbol{\lambda}_\star) \leq \frac{1}{T} \left(\frac{1}{2\gamma} \|(\mathbf{W}_0 \mathbf{X} - \mathbf{\Pi} \mathbf{S}_\mathfrak{h} + \mathbf{U}_0 - \gamma \boldsymbol{\lambda}_\star)\|^2 \right).$$

Proof sketch. The proof largely follows that of Theorem 2.1, except that in this case f is not convex. However, we will show that as long as the initialization $\mathbf{S}_0$ satisfies

$$\log |\det \mathbf{S}_0 \mathbf{S}_\mathfrak{h}^\dagger| \leq \max_{\mathbf{\Pi}, \mathbf{D}} \text{Tr } \mathbf{\Pi} \mathbf{D} \mathbf{S}_0 \mathbf{S}_\mathfrak{h}^\dagger - k,$$

then (11) is satisfied throughout the iterations. All we need is to prove the following lemma. $\square$

Lemma 3.2. *When running the L-ADMM iterations, if $\mathbf{S}_t$ satisfies*

$$\log |\det \mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger| \leq \text{Tr } \mathbf{\Pi} \mathbf{D} \mathbf{S}_t \mathbf{S}_\mathfrak{h}^\dagger - k,$$

for some permutation matrix $\mathbf{\Pi}$ and diagonal matrix $\mathbf{D}$ with ± 1 on the diagonal, then $\mathbf{S}_{t+1}$ also satisfies

$$\log |\det \mathbf{S}_{t+1} \mathbf{S}_\mathfrak{h}^\dagger| \leq \text{Tr } \mathbf{\Pi} \mathbf{D} \mathbf{S}_{t+1} \mathbf{S}_\mathfrak{h}^\dagger - k,$$

The proof is relegated to the supplementary.

4 EXPERIMENTS

We now provide some numerical experiments to showcase the effectiveness of the proposed L-ADMM for solving some special cases of (1), in particular BCA, SCA, NCA, and ACA. All the experiments are conducted in MATLAB on an iMac. We synthetically generate random problems. For $k = 20$ and $n = 1000$, we randomly generate the groundtruth factor matrices $\mathbf{S}^\mathfrak{h} \in \mathbb{R}^{k \times n}$ and $\mathbf{A}^\mathfrak{h} \in \mathbb{R}^{k \times k}$, and construct the data matrix $\mathbf{X} = \mathbf{A}^\mathfrak{h} \mathbf{S}^\mathfrak{h}$. Elements of $\mathbf{A}_\mathfrak{h}$ are independently drawn from a standard normal distribution, while those of $\mathbf{S}_\mathfrak{h}$ are generated as follows:

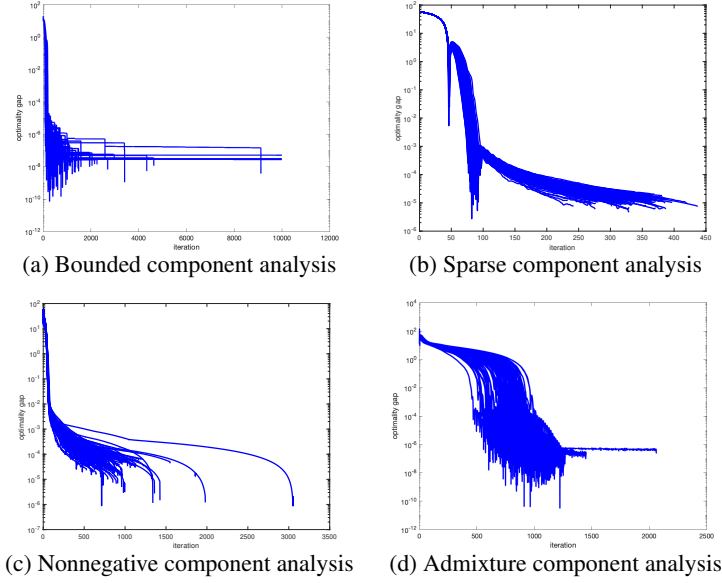


Figure 2: Convergence of L-ADMM in (2) for various latent component analyses models on 100 random trials.

- BCA: each S_{ij} has 50% chance to be ± 1 and 50% chance uniform in $[-1, 1]$.
- SCA: each S_{ij} has 50% chance to be 0 and 50% chance standard normal; then each row is rescaled to have unit ℓ_1 norm.
- NCA: each S_{ij} has 50% chance to be 0 and 50% chance exponential; then each row is rescaled to sum to one.
- ACA: each S_{ij} has 50% chance to be 0 and 50% chance exponential; then each column is rescaled to sum to one.

Prior works have shown that each of these generative models is identifiable with very high probability, despite n being not that big compared to the number of atoms k . Matrix $\mathbf{X}$ is used as input to the L-ADMM algorithm as described in (2). Although Problem (15) is nonconvex, as long as it is identifiable, we know the global optimum is attained at $\mathbf{W}_\star == \mathbf{A}_\dagger^\dagger$. As a result, $(1/2) \log \det \mathbf{A}_\dagger^\top \mathbf{A}_\dagger$ is the optimal value for Problem (15) as long as the model is identifiable, and we shall see whether the proposed algorithm is able to attain that optimal value. Inspired by the above convergence analysis, we check the optimality gap of the Lagrangian function values using the optimal dual variable $\mathbf{A}$

$$-(1/2) \log \det \mathbf{W}_t \mathbf{W}_t^\top + \text{Tr}(\mathbf{S}_t - \mathbf{W}_t \mathbf{X}) \mathbf{A}_\star - (1/2) \log \det \mathbf{A}_\dagger^\top \mathbf{A}_\dagger.$$

Furthermore, it is easy to show that an optimal $\mathbf{A}$ is $\mathbf{S}_\dagger^\dagger$. In this simulation with known groundtruth factors, we will use this to measure the optimality gap. The convergence behavior of 100 random trials of the L-ADMM are shown in Figure 2. Indeed, even though we are trying to solve a nonconvex problem (1), L-ADMM always converges to global optimum in our experiment. Each execution takes no more than a few seconds.

5 CONCLUSION

A general framework of volume optimization-based latent component analyses problems are studied, which includes many well-known unsupervised learning models such as dictionary learning, nonnegative matrix factorization, topic modeling, etc. An algorithm based on linearized ADMM (L-ADMM) is proposed, which admits simple update rules that are easy to implement in practice. Even though the problem is NP-hard, we show both in theory and in practice that the proposed algorithm is extremely effective at finding a global optimum.

REFERENCES

- Adi Ben-Israel. A volume associated with $m \times n$ matrices. *Linear algebra and its applications*, 167: 87–111, 1992.
- Adi Ben-Israel. The change-of-variables formula using matrix volume. *SIAM Journal on Matrix Analysis and Applications*, 21(1):300–312, 1999.
- Adi Ben-Israel. An application of the matrix volume in probability. *Linear Algebra and its Applications*, 321(1-3):9–25, 2000.
- David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.
- Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.
- John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1 -ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279, 2008.
- Xiao Fu, Wing-Kin Ma, Kejun Huang, and Nicholas D Sidiropoulos. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9):2306–2320, 2015.
- Xiao Fu, Kejun Huang, and Nicholas D Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.
- Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76:327–363, 2018.
- Bingsheng He, Li-Zhi Liao, Deren Han, and Hai Yang. A new inexact alternating directions method for monotone variational inequalities. *Mathematical Programming*, 92:103–118, 2002.
- Jingzhou Hu and Kejun Huang. Global identifiability of ℓ_1 -based dictionary learning via matrix volume optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023a.
- Jingzhou Hu and Kejun Huang. Identifiable bounded component analysis via minimum volume enclosing parallelotope. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023b.
- Jingzhou Hu and Kejun Huang. Complex bounded component analysis: Identifiability and algorithm. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6680–6684. IEEE, 2024.
- Kejun Huang, Nicholas D Sidiropoulos, and Ananthram Swami. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, 2013.
- A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley, 2004. ISBN 9780471464198. URL <https://books.google.com/books?id=96D0ypDwAkkC>.
- Chia-Hsiang Lin, Wing-Kin Ma, Wei-Chiang Li, Chong-Yung Chi, and ArulMurugan Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.
- Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. An extragradient-based alternating direction method for convex minimization. *Foundations of Computational Mathematics*, 17(1):35–59, 2017.

- Songtao Lu, Jason D Lee, Meisam Razaviyayn, and Mingyi Hong. Linearized admm converges to second-order stationary points for non-convex problems. *IEEE Transactions on Signal Processing*, 69:4859–4874, 2021.
- Wing-Kin Ma, José M Bioucas-Dias, Tsung-Han Chan, Nicolas Gillis, Paul Gader, Antonio J Plaza, ArulMurugan Ambikapathi, and Chong-Yung Chi. A signal processing perspective on hyperspectral unmixing: Insights from remote sensing. *IEEE Signal Processing Magazine*, 31(1):67–81, 2013.
- Michael K Ng, Fan Wang, and Xiaoming Yuan. Inexact alternating direction methods for image recovery. *SIAM Journal on Scientific Computing*, 33(4):1643–1668, 2011.
- Packer. Np-hardness of largest contained and smallest containing simplices for v-and h-polytopes. *Discrete & Computational Geometry*, 28:349–377, 2002.
- Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 2008.
- Yuchen Sun and Kejun Huang. Improved identifiability and sample complexity analysis of complete dictionary learning. In *2024 IEEE 13rd Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 1–5. IEEE, 2024.
- Yuchen Sun and Kejun Huang. Global identifiability of overcomplete dictionary learning via l1 and volume minimization. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Gokcan Tatli and Alper T Erdogan. Polytopic matrix factorization: Determinant maximization based criterion and identifiability. *IEEE Transactions on Signal Processing*, 69:5431–5447, 2021.