sensors

MDPI

Article

An Efficient Orthonormalization-Free Approach for Sparse Dictionary Learning and Dual Principal Component Pursuit

Xiaoyin Hu [1,2,*,†] **and Xin Liu** [1,2,3,†]

1 Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
2 School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
3 State Key Laboratory of Scientific and Engineering Computing, Beijing 100190, China; liuxin@lsec.cc.ac.cn
* Correspondence: hxy@amss.ac.cn
† These authors contributed equally to this work.

Received: 10 May 2020; Accepted: 25 May 2020; Published: 27 May 2020

Abstract: Sparse dictionary learning (SDL) is a classic representation learning method and has been widely used in data analysis. Recently, the ℓ_m-norm ($m \geq 3, m \in \mathbb{N}$) maximization has been proposed to solve SDL, which reshapes the problem to an optimization problem with orthogonality constraints. In this paper, we first propose an ℓ_m-norm maximization model for solving dual principal component pursuit (DPCP) based on the similarities between DPCP and SDL. Then, we propose a smooth unconstrained exact penalty model and show its equivalence with the ℓ_m-norm maximization model. Based on our penalty model, we develop an efficient first-order algorithm for solving our penalty model (PenNMF) and show its global convergence. Extensive experiments illustrate the high efficiency of PenNMF when compared with the other state-of-the-art algorithms on solving the ℓ_m-norm maximization with orthogonality constraints.

Keywords: dual principal component pursuit; orthogonality constraint; sparse dictionary learning; stiefel manifold

1. Introduction

In this paper, we focus on solving the optimization problem with orthogonality constraints:

$$\begin{aligned} \min_{W \in \mathbb{R}^{n\times p}} \quad & f(W) := -\frac{1}{m}\left\|W^\top Y\right\|_m^m \\ \text{s.t.} \quad & W^\top W = I_p, \end{aligned} \tag{1}$$

where W is the variable, $Y \in \mathbb{R}^{n\times N}$ is a given data matrix, and I_p denotes the identity matrix in $\mathbb{R}^{p\times p}$. Besides, the ℓ_m-norm is defined as $\|Y\|_m = \left[\sum_{i=1}^n \sum_{j=1}^N (Y_{ij})^m\right]^{\frac{1}{m}}$ with constant $m \in (2,4]$. For brevity, the orthogonality constraints $W^\top W = I_p$ in (1) can be expressed as $W \in \mathcal{S}_{n,p} := \{W \in \mathbb{R}^{n\times p} | W^\top W = I_p\}$. Here, $\mathcal{S}_{n,p}$ denotes the Stiefel manifold in real matrix space, and we call it the Stiefel manifold for simplicity in the rest of our paper.

The sparse dictionary learning (SDL) exploits the low-dimensional features within a set of unlabeled data, and therefore plays an important role in unsupervised representative learning. More specifically, given a data set $Y = [y_1, y_2, \ldots, y_N] \in \mathbb{R}^{n\times N}$ that contains N samples in $\mathbb{R}^n$, SDL aims to compute a full-rank matrix $D \in \mathbb{R}^{n\times p}$ named as dictionary, and an associated sparse representation $X = [x_1, \ldots, x_N]$ that satisfies

$$Y = DX, \tag{2}$$

or equivalently, find a $W = D^{\top -1}$ such that

$$X = W^\top Y. \tag{3}$$

As a result, the SDL can be solved by finding a $W \in \mathbb{R}^{n\times n}$, which leads to a sparse $W^\top Y$. Some existing works introduce the ℓ_0-norm or ℓ_1-norm penalty term to promote the underlying sparsity of $W^\top Y$ and present various algorithms for solving the consequent optimization models, see the work in [1–15] for instance. Interested readers are referred to a recent paper [16] and the references therein. However, the ℓ_1-norm minimization-based models are known to be sensitive with noise, and so far the existing approaches are not efficient enough for the purpose of solving real application problems which are often large-scale [17]. Consequently, a proper model with an efficient algorithm for SDL is desired, especially for the large-scale case.

Recently, an ℓ_4-norm maximization model is proposed in [17], which can recover the entire dictionary in a single run. This new formulation is motivated by the fact that maximizing a higher-order norm promotes spikiness and sparsity at the same time. The authors of [17] demonstrate that the global minimizers of ℓ_4-norm maximization with orthogonality constraints are very close to the true dictionary. Moreover, concaveness of the objective function in Equation (1) enables a fast fixed-point type algorithm, named matching, stretching, and projection (MSP). MSP achieves significant speedup compared with existing methods based on ℓ_0-norm or ℓ_1-norm penalty minimization. As maximizing any higher-order norm over a lower-order norm constraint leads to sparse and spiky solutions, Shen *et al.* [18] extend ℓ_4-norm maximization technique to a generalized l_m-norm maximization ($m \geq 3$). In addition, the authors propose a gradient projection method (GPM) for solving it with guaranteed global convergence.

However, both MSP and GPM invoke polar decomposition to keep the feasibility in each iteration. As illustrated in [19–21], orthonormalization lacks concurrency, which results in low scalability in column-wise parallel computing, particularly when the number of columns is large.

Several infeasible approaches have been developed to avoid orthonormalization. Gao *et al.* [19] propose the proximal linearized augmented Lagrangian method (PLAM) as well as its enhanced version, PCAL. Based on the merit function used in Gao *et al.* [19], Xiao *et al.* [21] propose an exact penalty model with a convex and compact auxiliary constraint, named PenC, for optimization problems with orthogonality constraints. The authors propose an approximated gradient method named PenCF for solving PenC and showed its global convergence and local convergence rate under mild conditions. The above-mentioned infeasible approaches do not require orthonormalization in each iteration. Numerical experiments illustrate the promising performance of these infeasible approaches with the existing state-of-the-art algorithms.

Although PCAL and PenCF avoid the orthonormalization process by taking infeasible steps, these approaches require additional constraints to restrict the sequence in a compact set in $\mathbb{R}^{n\times p}$, which can undermine their overall efficiency. Therefore, to develop an efficient algorithm on solving SDL, an infeasible model without constraints is desired.

Similar to the ℓ_1-norm penalty model for SDL, dual principal component pursuit (DPCP) aims to recover a tangent vector in $\mathbb{R}^n$ from samples $Y = [y_1, \ldots, y_n] \in \mathbb{R}^{n\times N}$ contaminated by outliers. Specifically, DPCP solves the following nonsmooth nonconvex optimization problem with a spherical constraint:

$$\begin{aligned} \min_{W\in\mathbb{R}^n} \quad & \left\|W^\top Y\right\|_1 \\ \text{s.t.} \quad & W^\top W = 1. \end{aligned} \tag{4}$$

Due to the ability on recovering an $n-1$ dimensional hyperplane from $\mathbb{R}^n$, DPCP has wide applications in 3D computer vision, such as detecting planar structures in 3D point clouds in KITTI dataset [22,23] and estimating relative poses in multiple-view [24].

Existing approaches [25–28] for solving convex problem (4) are not scalable and not competent in high dimensional cases [29]. On the other hand, the Random Sampling and Consensus (RANSAC) algorithm [30] has been one of the most popular methods in computer vision for the high relative dimension setting. RANSAC alternates between fitting a subspace to a randomly sampled minimal number of points ($n-1$ in the case of DPCP) and measuring the quality of selected subspace by using the number of data-points close to the subspace. In particular, as described in [29], RANSAC can be extremely effective when the probability of sampling outlier-free samples inside the allocated time budget is large. Recently, Tsakiris and Vidal [31] introduce Denoised-DPCP (DPCP-d) by minimizing $\left\|y - W^\top Y\right\|_{\mathrm{F}}^2 + \gamma \left\|y\right\|_1$ over the constraints $W^\top W = 1, y \in \mathbb{R}^N$. In the same paper, Tsakiris and Vidal [31] propose an Iteratively-Reweighted-Least-Squares algorithm (DPCP-IRLS) for solving the non-convex DPCP problem (4). The authors illustrate that DPCP-IRLS can successfully handle 30% to 50% of outliers and showed its high efficiency compared with RANSAC. In addition, Zhu *et al.* [32] propose a projected subgradient-based algorithm named DPCP-PSGM, which exhibits great efficiency on reconstructing road-plane in the KITTA dataset. There are also some approaches using smoothing techniques to approximate the ℓ_1-norm term such as Logcosh [8,33], Huber loss [34], pseudo-Huber [5], etc. Then, algorithms for minimizing a smooth objective function on a sphere can be applied. Nonetheless, these smoothing techniques often introduce approximation errors as the smooth objective functions usually lead to dense solutions. Qu *et al.* [35] and Sun *et al.* [8] propose a rounding step as postprocessing to achieve exact recovery [16] by solving a linear programming, which leads to addition computational cost.

The main difficulties in developing efficient algorithms are the nonsmoothness and nonconvexity in DPCP models. By observing the similarity between SDL and DPCP, we consider to adopt the ℓ_m-norm maximization to reformulate DPCP as a smooth optimization problem on sphere.

1.1. Contribution

In this paper, we first point out that the DPCP problem can be formulated as the ℓ_m-norm ($m \in (2,4]$) maximization (1) with $p = 1$. Therefore, both SDL and DPCP can be unified as a smooth optimization problem on the Stiefel manifold.

Motivated by PenC [21], we propose a novel penalty function as the following expression,

$$h(W) := f(W) - \frac{1}{2}\left\langle W^\top W - I_p, \Phi(W^\top \nabla f(W))\right\rangle + \frac{\beta}{6}\left\|W\right\|_{\mathrm{F}}^6 - \frac{\beta}{2}\left\|W\right\|_{\mathrm{F}}^2, \tag{5}$$

where $\beta > 0$ is the penalty-parameter and Φ is the operator that symmetrizes the square matrix, defined by $\Phi(M) = \frac{M+M^\top}{2}$. We show that $h(W)$ is bounded from below, then the convex compact constraint in PenC can be avoided. Therefore, we propose the following smooth unconstrained penalty model for ℓ_m-norm maximization (PenNM),

$$\min_{W \in \mathbb{R}^{n\times p}} \quad h(W). \tag{6}$$

We prove that Equation (6) with $m \in (2,4]$ is an exact penalty function of Equation (1) under some mild conditions. Moreover, when $p = 1$, we verify that PenNM does not introduce any first-order stationary point other than those of Equation (1) and $x = 0$. Based on the new exact penalty model, we propose an efficient orthonormalization-free first-order algorithm named PenNMF with no additional constraint. In PenNMF, we adopt an approximate gradient in each iterate instead of the exact one in which the second-order derivative of the original objective involves. The global convergence of PenNMF under mild conditions can be established.

The numerical experiments on synthetic and real imaginary data demonstrate that PenNMF outperforms PenCF and MSP/GPM in solving SDL, especially in large-scale cases. As an infeasible method, PenNMF shows superior performance when compared with MSP and GPM, which invoke an orthonormalization process to keep the feasibility. Moreover, when compared with

PenCF, PenNMF also shows better performance, implying the benefits of avoiding the constraints in PenC. In our numerical experiments on DPCP, our proposed model (1) with $p = 1$ shows comparable accuracy with ℓ_1-norm based penalty model (4) on solving road-plane recovery in KITTA dataset. In some test examples, (1) can have even better accuracy than (4). Besides, PenNMF takes less CPU time while achieving comparable accuracy in reconstructing road-plane in KITTA dataset when compared with other state-of-the-art algorithms such as DPCP-PSGM and DPCP-d.

1.2. Notations and Terminologies

Norms: In this paper, $\|\cdot\|_m$ denotes the element-wise m-th norm of a vector or matrix, i.e., $\|A\|_m = \left(\sum_{i=1}\sum_{j=1}|A_{ij}|^m\right)^{1/m}$. Besides, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $\|\cdot\|$ denotes the 2-th operator norm, i.e., $\|A\|$ equals the maximum singular value of A. Besides, we denote $\sigma_{\min}(A)$ as the smallest singular value of a given matrix A. The operator $A \circ B$ stands for the Hadamard product of matrices A and B with the same size. $|A|$ and $A^{\circ l}$ represent the component-wise absolute value and l-th power of matrix A, respectively. Besides, for two symmetric matrices A and B, $A \succeq B$ denotes that $A - B$ is semi-positive definite, and $A \succ B$ denotes that $A - B$ is positive definite.

Optimality Condition: W is a first-order stationary point of (1) if and only if

$$\begin{cases} (I_n - WW^\top)\nabla f(W) &= 0; \\ W^\top \nabla f(W) &= \nabla f(W)^\top W; \\ W^\top W &= I_p. \end{cases} \tag{7}$$

Besides, W is a first-order stationary point of PenNM if and only if $\nabla h(W) = 0$.

2. Model Description

In this section, we first discuss how to reformulate DPCP as an ℓ_m-norm maximization with orthognoality constraints. To construct a orthonormalization-free algorithm, we minimize $h(W)$ rather than directly solve (1). As an unconstrained penalty problem for (1), the model (6) may introduce additional infeasible first-order stationary points. Therefore, in this section, we characterize the equivalence between (1) and (6) to provide theoretical guarantees for our approach.

2.1. ℓ_m-Norm Maximization for DPCP Problems

Based on the fact that maximization of a higher-order norm promotes spikiness and sparsity, we maximize the ℓ_m-norm of $\hat{W}^\top Y$ over the constraint $\hat{W}^\top Y Y^\top \hat{W} = 1$. The model can be expressed as

$$\begin{aligned} \min_{\hat{W}\in\mathbb{R}^n} \quad & -\frac{1}{m}\left\|\hat{W}^\top Y\right\|_m^m \\ \text{s.t.} \quad & \hat{W}^\top Y Y^\top \hat{W} = 1. \end{aligned}$$

Although with different constraints to (1), (4) can be reshaped to the formulation of (1). Let $Y = RZ$ be the rank-revealing QR decomposition of Y, where $Z \in \mathbb{R}^{n\times N}$ is an orthogonal matrix and $R \in \mathbb{R}^{n\times n}$ is an upper-triangular matrix, and denote $W = R^{-T}\hat{W}$, then the optimization model can be reshaped as

$$\begin{aligned} \min_{W\in\mathbb{R}^n} \quad & -\frac{1}{m}\left\|W^\top Z\right\|_m^m \\ \text{s.t.} \quad & W^\top W = 1. \end{aligned} \tag{8}$$

Clearly, problem (8) is a special case of (1) with $p = 1$. Moreover, suppose W^* is a global minimizer of (8), the solution for DPCP problem can be recovered by $\hat{W}^* = R^{-T}W^*$. The detailed framework for solving DPCP by ℓ_m-norm maximization is presented in Algorithm 1.

Algorithm 1 Framework for Solving DPCP by ℓ_m-Norm Maximization.

Require: Data matrix $Y \in \mathbb{R}^{n\times N}$
1: Perform QR-factorization for Y. Namely, $Y = RZ$ where R is upper-triangular matrix and $Z \in \mathbb{R}^{n\times N}$ is orthogonal matrix;
2: Compute the solution $\hat{W}$ for (1);
3: Return $W^* = R^{-T}\hat{W}$

2.2. Equivalence

In this subsection, we first derive the expression for $\nabla f(W)$ and $\nabla h(W)$.

Proposition 1. *The gradient and the Hessian of $f(W)$ can be expressed as*

$$\nabla f(W) = Y\left[|Y^\top W|^{\circ(m-1)} \circ \operatorname{sign}((Y^\top W))\right];$$
$$\nabla^2 f(W)[D] = (m-1)Y\left[(Y^\top W)^{\circ(m-2)} \circ (Y^\top D)\right],$$

respectively. Moreover, the gradient of $h(W)$ can be formulated as

$$\begin{aligned}\nabla h(W) =&\nabla f(W)\left(\frac{3}{2}I_p - \frac{1}{2}W^\top W\right) - W\Phi(W^\top \nabla f(W)) - \frac{1}{2}\nabla^2 f(W)[W(W^\top W - I_p)]\\ &+ 2\beta WW^\top W(W^\top W - I_p) + \beta W(W^\top WW^\top W - I_p).\end{aligned}$$

Proof. From the work in [17] we have $\nabla f(W) = Y\left[|Y^\top W|^{\circ(m-1)} \circ sign(Y^\top W)\right]$. Based on the expression for $\nabla f(W)$, the Hessian of f can be expressed as $\nabla^2 f(W)[D] = (m-1)Y\left[(Y^\top D) \circ (Y^\top W)^{\circ(m-2)}\right]$. As a result, $\nabla^2 f(W)[W(W^\top W - I_p)] = (m-1)Y(Y^\top W)^{\circ(m-1)}Y(W^\top W - I_p)$.

Therefore, based on ([21], Equation 2.8), the gradient of $h(W)$ can be formulated as

$$\begin{aligned}\nabla h(W) =&\nabla f(W)\left(\frac{3}{2}I_p - \frac{1}{2}W^\top W\right) - W\Phi(W^\top \nabla f(W)) - \frac{1}{2}\nabla^2 f(W)[W(W^\top W - I_p)]\\ &+ 2\beta WW^\top W(W^\top W - I_p) + \beta W(W^\top WW^\top W - I_p)\end{aligned}$$

□

With the expression for $\nabla h(W)$, we can establish the equivalence between (1) and our proposed model, (6). The equivalence is illustrated in Theorem 4, and the main body of the proofs is presented in Appendix A.

Theorem 2. (First-order equivalence) *Suppose $\beta \geq (4m+8)\|Y\|_{\mathrm{F}}^m$ and $\tilde{W}$ is a first-order stationary point of (6), then either $\tilde{W}^\top\tilde{W} = I_p$ holds, which further implies that $\tilde{W}$ is a first-order stationary point of problem (1), or the inequality $\sigma_{\min}(\tilde{W}^\top\tilde{W}) \leq \frac{(2m+4)\|Y\|_{\mathrm{F}}^m}{\beta}$ holds.*

Theorem 2 characterizes the relationship between the first-order stationary points of (1) and those of (6). Namely, the penalty model only yields the first-order stationary points other than those of the original model (1) far away from the Stiefel manifold. When $p = 1$, we can derive a stronger result on those additional first-order stationary points produced by the penalty model in Corollary 3.

Corollary 3. (Stronger first-order equivalence for $p = 1$) *Suppose $p = 1$ in (1), $\beta \geq (4m+8)\|Y\|_{\mathrm{F}}^m$, and $\tilde{W}$ is a first-order stationary point of (6), then either $\tilde{W}^\top\tilde{W} = I_p$ holds, which further implies that $\tilde{W}$ is a first-order stationary point of problem (1), or $\tilde{W} = 0$.*

Theorem 2 characterizes the equivalence between (1) and (6) in the sense that all the infeasible first-order stationary points of (6) is relatively far away from the constraint $W^\top W = I_p$. Besides, Corollary 3 shows that when $p = 1$, the only infeasible first-order stationary point of (6) is 0. Therefore, when we achieve a solution near the constraint $W^\top W = I_p$ by solving (1), we can conclude that W is a first-order stationary point of (1). Instead of directly solving (1), we can compute the first-order stationary point of (6) and thus avoid intensive orthonormalization in the computation.

3. Algorithm

3.1. Global Convergence

In this section, we focus on developing an infeasible approach for solving (6). The calculation of the gradient of $h(W)$ is involved with the second-order derivative, which is typically even more expensive than the iterations in MSP/GPM. Therefore, we consider to solve (6) by an approximated gradient descent algorithm. Let $D(W) := \nabla f(W) - W\Phi(W^\top \nabla f(W)) + \beta W(W^\top W W^\top W - I_p)$ be the approximation for the gradient of $h(W)$, we present the detailed algorithm as Algorithm 2.

Algorithm 2 First-Order Method for Solving (6). (PenNMF)

Require: $f : \mathbb{R}^{n\times p} \mapsto \mathbb{R}$, $\beta > 0$;

1: Randomly choose W^0 satisfies ${W^0}^\top W^0 = I_p$, set $k = 0$;
2: **while** not terminate **do**
3: Compute inexact gradient

$$D(W^k) = \nabla f(W^k) - W^k \Phi({W^k}^\top \nabla f(W^k)) + \beta W^k({W^k}^\top W^k {W^k}^\top W^k - I_p);$$

4: Compute stepsize η^k;
5: $W^{k+1} = W^k - \eta^k D(W^k)$;
6: $k++$;
7: **end while**
8: Return W^k

Next, we establish the convergence of PenNMFin Theorem 4, which illustrates the global convergence and worst-case convergence rate of PenNMF under mild conditions. The main body of the proof is presented in Appendix B.

Theorem 4. (Global convergence) *Suppose* $\delta \in \left(0, \frac{1}{3}\right]$ *and* $\beta \geq \max\left\{228m\left\|Y\right\|_{\mathrm{F}}^m, \frac{32}{\delta}\left\|Y\right\|_{\mathrm{F}}^m\right\}$. *Let* $\{W^k\}$ *be the iterate sequence generated by PenNMF, starting from any initial point* W^0 *satisfying* $||{W^0}^\top W^0 - I_p||_{\mathrm{F}}^2 \leq \frac{1}{8}\delta$, *and the stepsize* $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, *where* $\bar{\eta} \leq \frac{1}{2M_1}$. *Then,* W^k *weakly converges to a first-order stationary point of* (1). *Moreover, for any* $k = 1, 2, \cdots$, *the convergence rate of PenNMF can be estimated by*

$$\min_{0\leq i\leq k}\left\|D(W^i)\right\|_{\mathrm{F}} \leq \sqrt{\frac{8m\left\|Y\right\|_{\mathrm{F}}^m + 2\beta\delta}{\bar{\eta}(k+1)}}. \tag{9}$$

3.2. Some Practical Settings

As illustrated in Algorithm 2, the hyperparameters in PenNMF are the penalty parameter β and stepsize η^k. In the theoretical analysis for PenNMF, the upper bound of η^k adopted in Theorem 4 is too restrictive in practice. There are many adaptive stepsize for first-order algorithms, and here we consider the Barzilai–Borwein (BB) stepsize [36],

$$\eta_{BB1,k} = \frac{\langle S_k, Y_k\rangle}{\langle Y_k, Y_k\rangle}, \quad \eta_{BB2,k} = \frac{\langle S_k, S_k\rangle}{\langle S_k, Y_k\rangle}, \tag{10}$$

and alternating Barzilai–Borwein (ABB) stepsize [37],

$$\eta_{ABB}^k = \begin{cases} \eta_{BB1}^k & mod(k,2) = 1 \\ \eta_{BB2}^k & mod(k,2) = 0, \end{cases} \tag{11}$$

where and $S_k = W^k - W^{k-1}$, $Y_k = \nabla h(W^k) - \nabla h(W^{k-1})$. We suggest to choose the stepsize η^k as ABB stepsize in PenNMF, and we test PenNMF with ABB stepsize in our numerical experiments.

Another parameter is β, which controls the smooth penalty term in $h(W)$. Similarly, the lower-bound for β in Theorem 4 is too large to be practical. In our numerical examples, we uses the constant $s := \left\|\nabla f(W^0)\right\|_{\mathrm{F}}$, which is an approximation to $\left\|\nabla^2 f(W^0)\right\|_{\mathrm{F}}$, as an upper-bound for β. According to the work in [21], we suggest to choose the penalty parameter by $\beta = 0.01\left\|\nabla^2 f(W^0)\right\|_{\mathrm{F}}$.

Additionally, to achieve high accuracy in feasibility, we perform the polar factorization to the final solution generated by PenCF and PenNMF as the default postprocess. More precisely, when we compute the final solution W^k by PenNMF, we can compute its rank-revealing singular-value decomposition $W^k = U^k \Sigma^k {V^k}^\top$ and return $\hat{W}^k := U^k {V^k}^\top$. Using the same proof techniques in [21], our postprocess leads to decrease in feasibility as well as the functional value. Moreover, the numerical experiments in [19] show that the introduced orthonormalization process results in little changes in $\nabla h(W)$. Therefore, we suggest to perform the described postprocess for PenNMF.

4. Numerical Examples

In this section, we present our preliminary numerical examples. We compare our algorithm with some state-of-the-art algorithms on SDL and DPCP problems, which are formulated as (1) and (8), respectively. Then, we observe the performance of our algorithm under different selections of parameters, and then choose the default setting. All the numerical experiments in this section are tested on an Intel(R) Core(R) Silver 4110 CPU @ 2.1 GHz, with 32 cores and 394 GB of memory running under Ubuntu 18.04 and MATLAB R2018a.

4.1. Numerical Results on Sparse Dictionary Earning

In this subsection, we mainly compare the numerical performance of PenNMF with some state-of-the-art algorithms on SDL. As illustrated in Table 2 in [17], MSP is significantly faster than the Riemannian subgradient [3] and Riemannian trust-region method [8]. Therefore, to have a better illustration on the performance of PenNMF, we compare PenNMF with state-of-the-art algorithms on solving (1), which is a smooth optimization problem with orthogonality constraints. We first select two state-of-the-art algorithms on solving optimization problems with orthogonality constraints. One is Manopt [38,39], a projection-based feasible method. In our numerical test, we choose nonlinear conjugate gradient with inexact linear-search strategy to accelerate Manopt. Another one is PenCF [21], which is an infeasible approach for optimization problems with orthogonality constraints. In our algorithms we choose to apply Alternating Bzarzilar–Borwein stepsize to accelerate PenNMF, and uses all parameters as default setting described in [21]. Besides, we test the MSP algorithm [17] and GPM algorithm [18]. It is worth to mention that when $m = 4$, the MSP and GPM are actually the same. According to the numerical examples in [18], $m = 3$ has better recovery quality than the case $m = 4$. Therefore, in our numerical experiments, we test the mentioned algorithms on the case where $m = 3$.

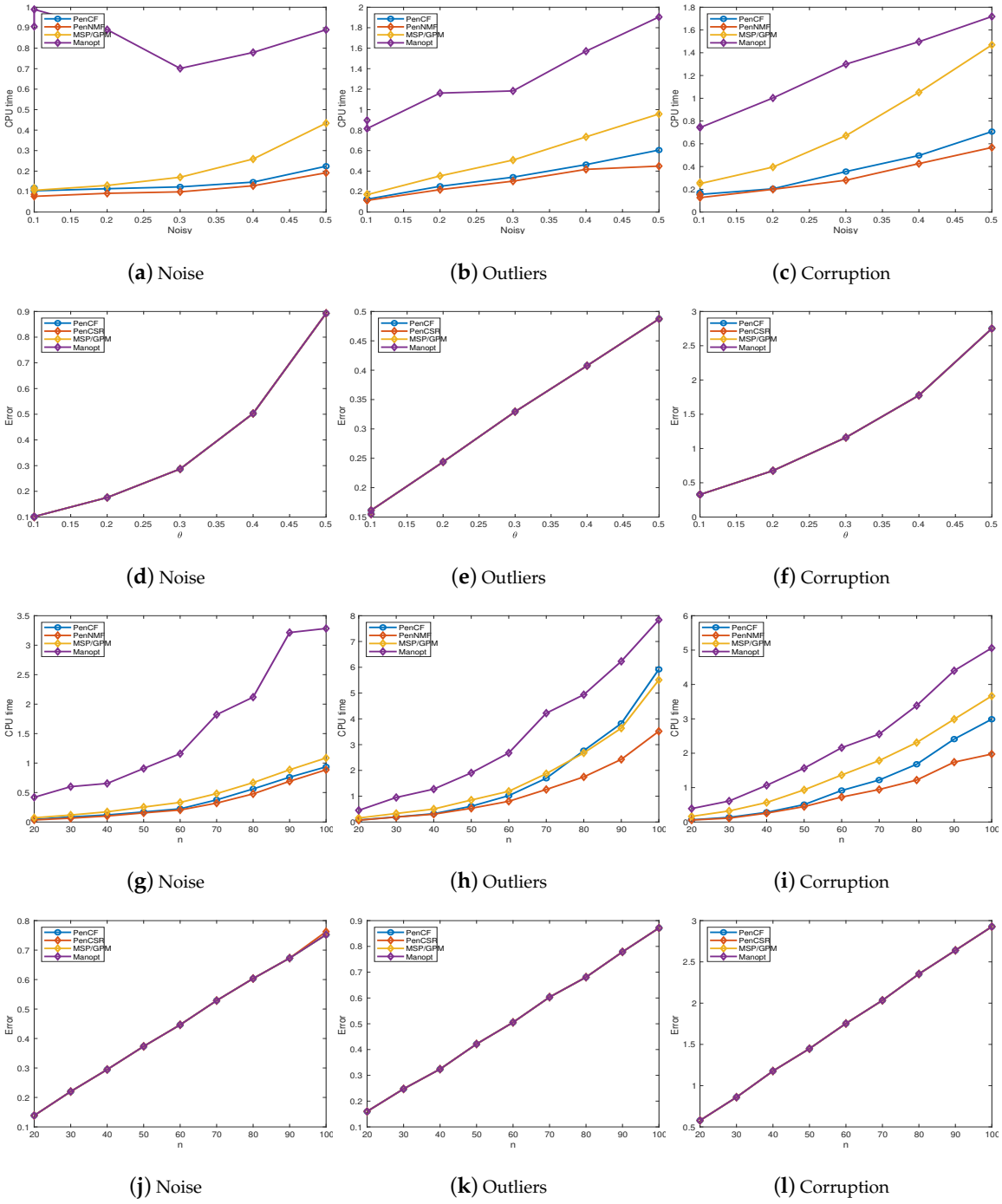
The stopping criteria for Manopt , MSP/GPM is $\left\|\nabla f(W^k) - W^k \Lambda(W^k)\right\|_{\mathrm{F}} \leq 10^{-2}$, while the stopping criteria for PenCF and PenNMF is $\left\|\nabla h(W^k)\right\|_{\mathrm{F}} \leq 10^{-2}$. Besides, the max iteration for all compared algorithms is set as 200.

In all test examples, we randomly generate the sparse representation X by $X^* = \mathrm{randn(n, N)}$. $*\,(\mathrm{randn(n, N)} < 0.3)$ and the dictionary W^* by randomly selecting a point on Stiefel manifold. Then, the original data matrix Y is constructed by $Y_o = {W^*}^\top X^*$. To test the performance of

all compared algorithms, we add different types of noise to Y_o. We first fix the level of noise $\theta = 0.3$ and choose n from 20 to 100. Then, we test the performance of compared algorithms with different types of noisy while fix $n = 50$. In our numerical tests, the "Noise" denotes the Gaussian Noise, where Y is constructed by $Y = Y_o + \theta \cdot \text{randn}(n, N)$. Besides, the term "Outliers" denotes the Gaussian outliers, where outliers $=$ randn$(n, \text{round}(\theta m))$; $Y =$ cat$(2, Y_o, \text{outliers})$;. Additionally, the term "Corruption" refers to the Gaussian corruption to Y_o, which is achieved by rademacher $= (\text{rand}(n, m) < 0.5) * 2 - 1; Y = Y_o + (\text{rand}(n, m) < \theta). *$ rademacher. Besides, the term 'CPU time' denotes the averaged run-time, while the term 'Error' denotes the $1 - \left\|\hat{W}^\top W^*\right\|_4^4$, where $\hat{W}$ denotes the final output of all the compared algorithm.

The numerical results are listed in Figure 1. From Figure 1d–f, j–l we conclude that all these compared algorithms achieve almost the same accuracy in all the cases. Besides, for Gaussian noise, the performance of PenNMF is comparable to MSP/GPM algorithm and outperforms Manopt. Moreover, with Gasuuain outliers and Gaussian corruption, the performance of PenNMF is better than PenCF, MSP/GPM, and Manopt. One possible explanation is that for Manopt invokes computing the Riemannian gradient, line-search in each iteration, resulting in higher computational complexity than MSP/GPM. Besides, the infeasible approaches overcome the bottleneck in the orthonormalization process in Manopt and MSP/GPM, and thus achieve comparable performance to MSP/GPM. Additionally, PenCF solves a constrained model by taking approximated gradient descent steps, while in PenNMF the model is an unconstrained one. The absence of constraint helps to improve the performance of PenNMF.

Besides testing on synthetic datasets, we also perform extensive experiments to verify the performance of PenNMF on real imagery data. A classic application of dictionary learning involves learning sparse representations of image patches [40]. In this paper, we extend the experiments in [17] to learn patches from grayscale and color images. Based on the 512×512 grayscale image "Barbara", we construct the clean data matrix Y_o by vectorizing each 16×16 patches from it. Then, we use the same approach to construct the clean data matrix Y from 512×512 grayscale images "Boat" and "Lena", together with a 256×256 grayscale image "House". In "Barbara", "Boat", and "Lena", the clean data matrix $Y \in \mathbb{R}^{256 \times 247{,}009}$, and the data matrix from "House" satisfies $Y \in \mathbb{R}^{256 \times 58{,}081}$. Besides, we construct the matrix $Y \in \mathbb{R}^{192 \times 62{,}001}$ by vectorizing the $8 \times 8 \times 3$ patches from the 256×256 RGB image "Duck". In such setting, all the compared algorithms recover the dictionary for all three channels simultaneously rather than learn them once for each channel in "Duck". Such approach is aslo applied to generate the data matrix in $\mathbb{R}^{192 \times 146{,}633}$ from 338×450 RGB image "Chateau". We run MSP/GPM, PenNMF, PenCF, and Manopt with $m = 3$ to compute the dictionary from $Y = Y_o + \theta \cdot \text{randn}(n, N)$ with different level of noise, where Y_o is generated in the same manner as our first numerical experiment and has the same size as these patched figures. The numerical results are presented in Figure 2 and Figure A1. In all experiments, PenNMF takes less time than PenCF, MSP/GPM, and Manopt, which further illustrate the high efficiency of PenNMF in tackling the real imagery data, especially in the large-scale case.



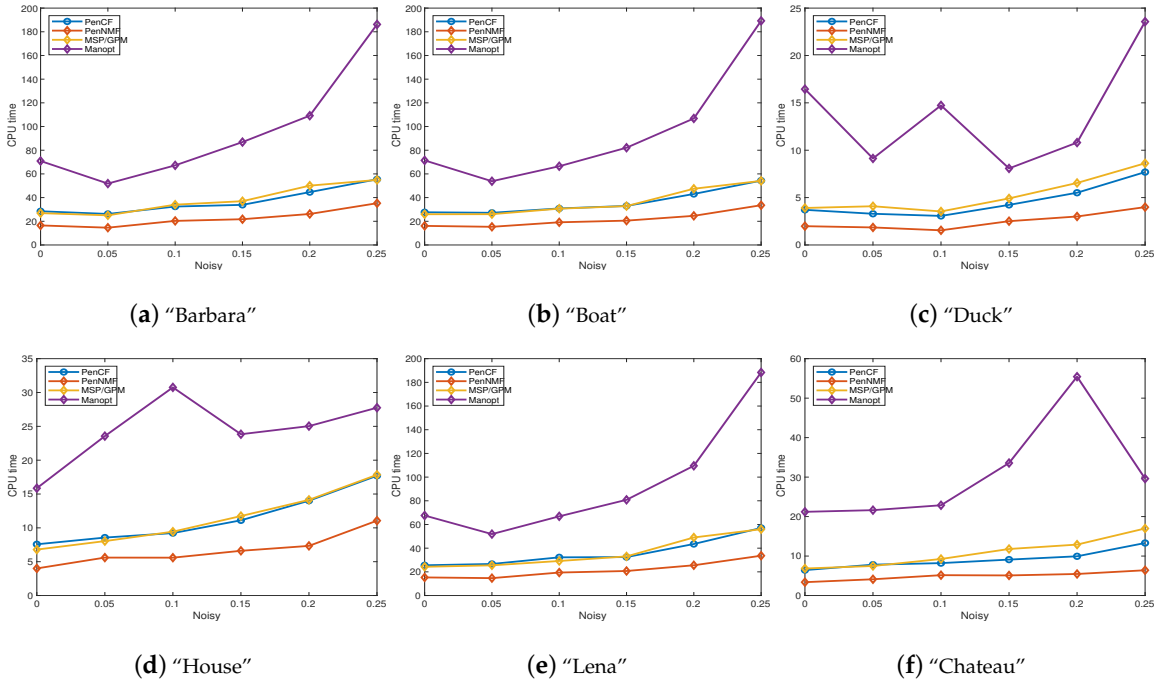
(**a**) Noise (**b**) Outliers (**c**) Corruption

(**d**) Noise (**e**) Outliers (**f**) Corruption

(**g**) Noise (**h**) Outliers (**i**) Corruption

(**j**) Noise (**k**) Outliers (**l**) Corruption

Figure 1. A detailed comparison among MSP, Manopt, PenCF, and PenNMF. (**a**)–(**c**) a comparison with different level of noisy on CPU time; (**d**)–(**f**) a comparison with different level of noisy on errors; (**g**)–(**i**) a comparison with different n on their CPU time; (**j**)–(**l**) a comparison with different n on their errors. The errors are evaluated by $1 - \left\|\hat{W}^\top W^*\right\|_4^4$, where $\hat{W}$ denotes the final output of all the compared algorithm.



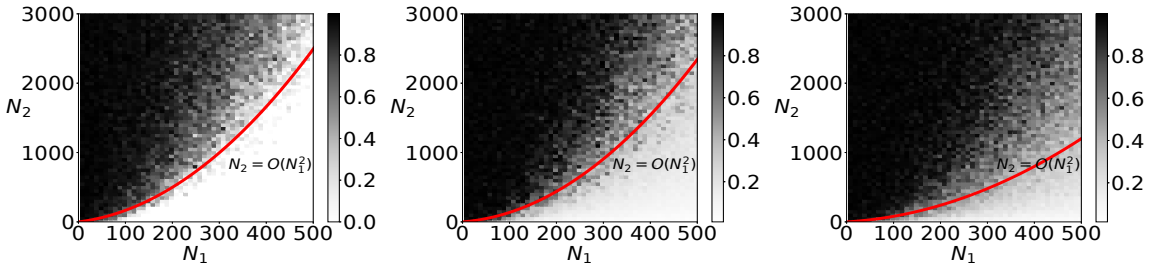
(**a**) "Barbara" (**b**) "Boat" (**c**) "Duck"

(**d**) "House" (**e**) "Lena" (**f**) "Chateau"

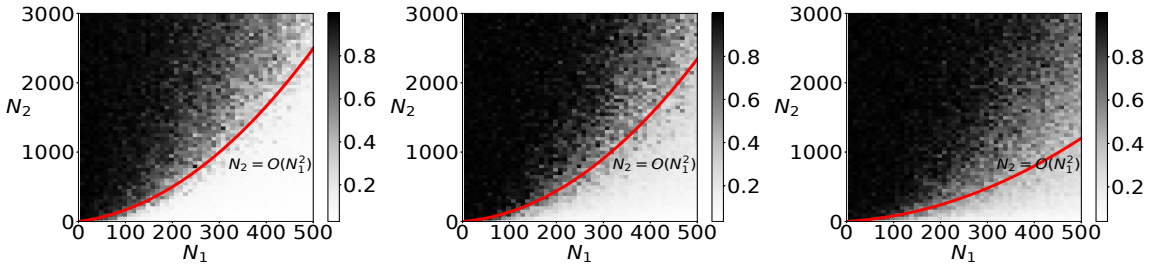
Figure 2. The CPU time of PenCF, PenNMF, MSP/GPM, and Manopt on computing the dictionary. (**a**) Barbara, $Y \in \mathbb{R}^{256\times 247,009}$; (**b**) Boat, $Y \in \mathbb{R}^{256\times 247,009}$; (**c**) Duck, $Y \in \mathbb{R}^{192\times 62,001}$; (**d**) House, $Y \in \mathbb{R}^{256\times 58,081}$; (**e**) Lena, $Y \in \mathbb{R}^{256\times 247,009}$; (**f**) Chateau, $Y \in \mathbb{R}^{192\times 146,633}$.

4.2. Dual Principal Component Pursuit

In this subsection, we first verify the recovery property of our proposed model (8), which is a special case of (1) by fixing $p = 1$. We first compare the distance between global minimizer of (8) and the ground-truth for DPCP problem. We first fix $n = 30$ and randomly select $W^* \in \mathbb{R}^n$. Then, we randomly generate N_1 inliers in the hyperplane whose normal vector is W^*. Besides, we randomly generate N_2 outliers in $\mathbb{R}^n$ following Gaussian distribution. Additionally, the data is corrupted by Gaussian noise by adding $\frac{\theta}{\sqrt{n}} \cdot \text{randn}(n, N)$ to Y. Then, we normalize each sample in Y. The range of N_1 is $[10, 500]$, whereas the range of N_2 is $[10, 3000]$. We run each test problem for 5 instances. Moreover, in each instance, we run DPCP-PSGM to solve (4) and PenNMF to solve (8) with $m = 3$ and 4, and get the solution $\tilde{W}$ for each model. We plot the principal angle between $\tilde{W}$ and W^* in Figure 3. From Figure 3a and 3d we can conclude that (4) can tolerate $\mathcal{O}(N_1^2)$ outliers while achieve exact recovery, which coincides the theoretical results presented in [32]. For model (8), numerical experiments do not show the exact recovery ability of (8) for $m = 3$ and 4. However, with some tolerance on the principal angle, we also observe that (4) can tolerate $\mathcal{O}(N_1^2)$ outliers. Moreover, we conclude that with $m = 3$, (8) has better ability to recover the normal vector than $m = 4$. As a result, in the rest of this subsection, we only test (8) with $m = 3$. In addition, we analyze the number of successfully recovered instances, where the $\sqrt{1 - \langle \tilde{W}, W^* \rangle^2}$ is less than 0.1 or 0.2. The results are presented in Figure 4. From Figure 4, we can conclude that, with tolerance on the errors, the ℓ_m-norm maximization model can successfully recover the normal vector. Moreover, in model (8), $m = 3$ has better performance than $m = 4$, which coincides with the numerical experiments in [18]. Therefore, when applying ℓ_m-norm maximization model to solving the DPCP problems, we suggest to choose $m = 3$ in (8).

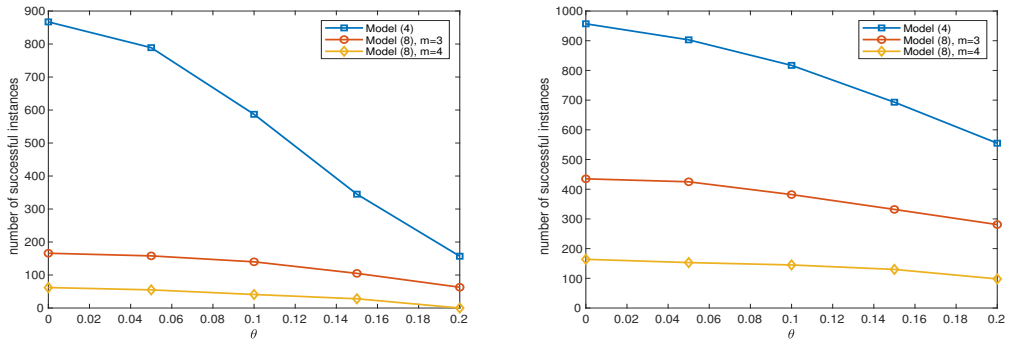


(**a**) Model (4), $\theta = 0$ (**b**) Model (8), $m = 3$, $\theta = 0$ (**c**) Model (8), $m = 4$, $\theta = 0$



(**d**) Model (4), $\theta = 0.1$ (**e**) Model (8), $m = 3$, $\theta = 0.1$ (**f**) Model (8), $m = 4$, $\theta = 0.1$

Figure 3. A comparison between the models (8) and (4) on the average recovery error $\sqrt{1 - \langle \tilde{W}, W^* \rangle^2}$ of 5 random trials. (**a**)–(**c**) average recovery errors with $\theta = 0$; (**d**)–(**g**) average recovery errors with $\theta = 0.1$.



(**a**) tolerance = 0.1 (**b**) tolerance = 0.2

Figure 4. A comparison on the number of successfully recovered instances on the different level of noise. (**a**) $\sqrt{1 - \langle \tilde{W}, W^* \rangle^2}$ is less than 0.1; (**b**) $\sqrt{1 - \langle \tilde{W}, W^* \rangle^2}$ is less than 0.2.

In the rest of this subsection, we test the numerical performance of PenNMF on solving DPCP problem, which plays an important role in autonomous driving applications. DPCP is applied to recover the road-plane, which can be regarded as inliers, from the 3d point clouds in KITTA dataset [22], which is recorded from a moving platform while driving in and around Karlsruhe, Germany. This dataset consists of image data together with corresponding 3D points collected by a rotating 3D laser scanner [32]. Moreover, DPCP only uses the 3D point clouds with the objective of determining the 3D points that lie on the road plane (inliers) and those off that plane (outliers): Given a 3D point cloud of a road scene, the DPCP problem focuses on reconstructing an affine plane $\{x \in \mathbb{R}^3 | a^\top x - b = 0\}$ as a representation for the road. Equivalently, this task can be converted to a linear subspace learning problem by embedding the affine plane into the linear hyperplane $H \subseteq \mathbb{R}^4$ with normal vector $\tilde{b} = [a, -b]$, through the mapping $x \to [x, 1]$ [29]. We use the experimental set-up

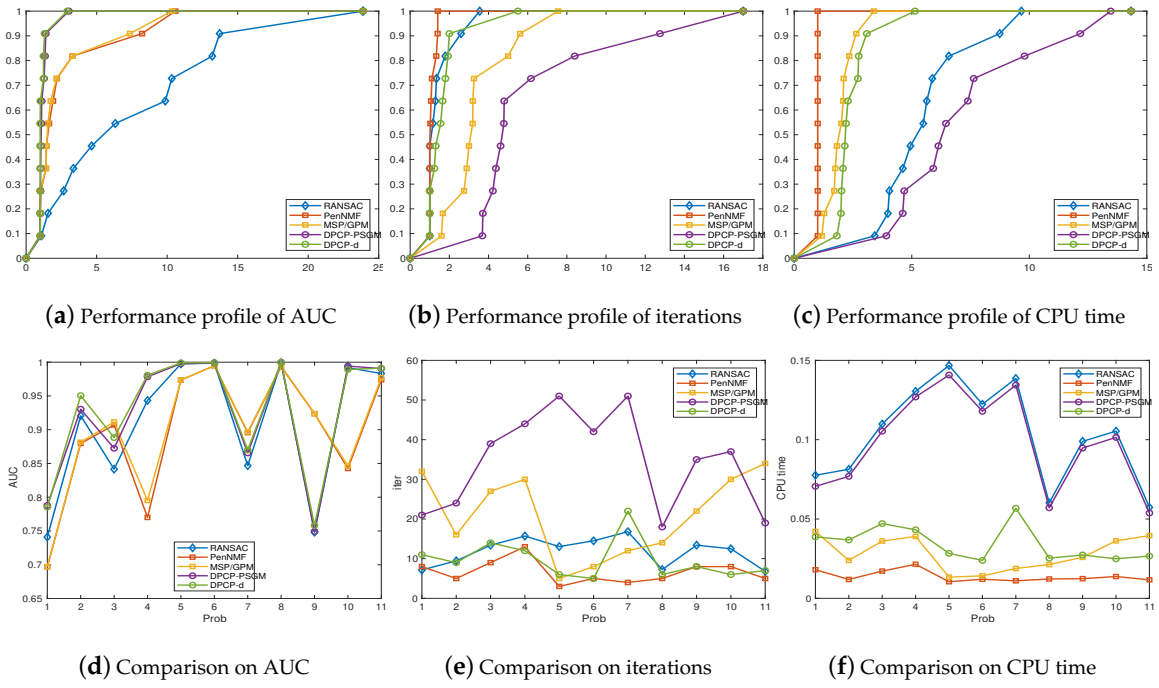
in [29,32] to further compare Equations (4) and (8), RANSAC, and other alternative methods in the task of 3D road plane detection in KITTA dataset. Each point cloud contains over 10^5 samples with approximately 50% outliers. Besides, the samples are homogenized and normalized to unit ℓ_2-norm.

We use 11 frames annotated in [29,32] from KITTA dataset. We compare DPCP-PSGM [29], DPCP-IRLS, and DPCP-d [31], which focus on solving the ℓ_1-norm minimization model (4). Besides, we test RANSAC and $\ell_{2,1}$-RPCA [25]. Additionally, we test PenNMF and MSP/GPM on solving our proposed model (8), which is a special case of (1). For DPCP-PSGM, DPCP-d, DPCP-IRLS, and $\ell_{2,1}$-RPCA, all parameters are set by following the suggestions in [32].

Figure 5 illustrates the numerical performance of all the compared algorithms. We present the numerical results in Figure 5d–f. Moreover, we draw the performance profiles proposed by Dolan and Moré [41] in Figure 5a–c to present an illustrative comparison on the performance of all compared algorithms. The performance profiles can be regarded as distribution functions for a performance metric for benchmarking and comparing optimization algorithms. Besides, we draw the recovery results of frames 328 and 441 in KITTA-CITY-71, which is presented in Figure 6. Here the term "AUC" denotes the area under the AUC curve, and "iterations" denotes the total iterations taken by these compared algorithms. Besides, "Prob" in Figure 5d–f denotes the indexes of tested frames, which are presented in Table 1.

Table 1. The testing instances and their corresponding frames in KITTA dataset.

Dataset	**KITTI-CITY-71**				**KITTI-CITY-5**					**KITTI-CITY-48**	
Frame id.	221	328	441	881	1	45	120	137	153	0	21
Test id.	1	2	3	4	5	6	7	8	9	10	11



(**a**) Performance profile of AUC (**b**) Performance profile of iterations (**c**) Performance profile of CPU time

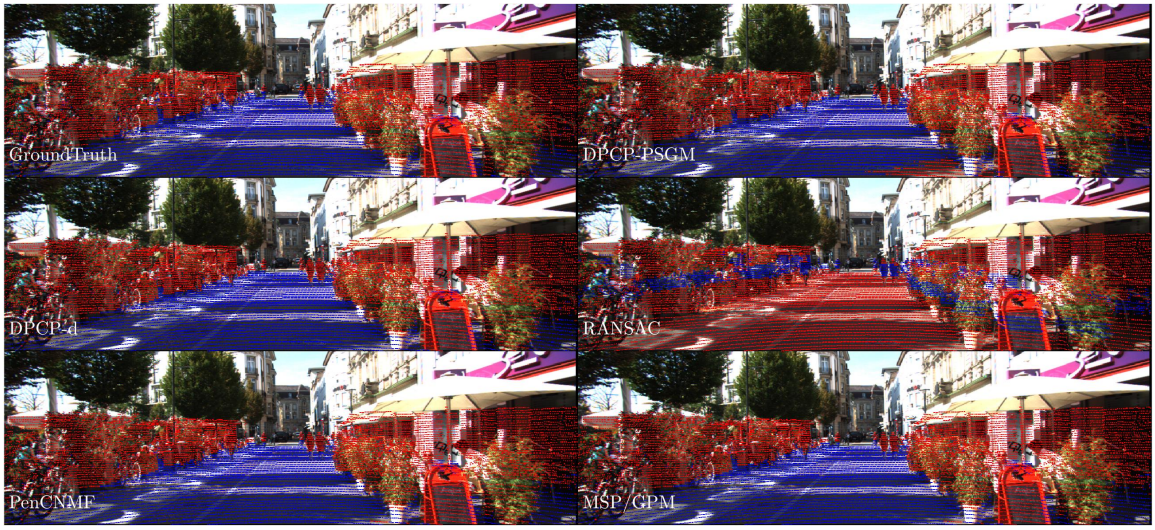
(**d**) Comparison on AUC (**e**) Comparison on iterations (**f**) Comparison on CPU time

Figure 5. A comparison between PenNMF , MSP, DPCP-PSGM, DPCP-D, and Random Sampling and Consensus (RANSAC). (**a**)–(**c**) performance profile [41] of AUC, iterations and CPU time; (**d**)–(**f**) the numerical results of AUC, iterations and CPU time.

From Figure 5a, we can conclude that PenNMF and MSP/GPM successfully recover the hyperplanes with comparable accuracy. Moreover, in problems 3, 7, and 9, PenNMF and MSP produce better classification accuracy than other approaches. Besides, in the aspect of CPU time,

PenNMF and MSP cost much less time than other compared algorithms in most cases. Moreover, from Figure 5c, we can conclude that PenNMF takes less time than MSP as well as other compared algorithms in almost all the cases. As a result, we can conclude that our proposed model (1) is easy to be solved and PenNMF shows better efficiency than MSP in our test examples.

(**a**) Frame 328 of KITTI-CITY-71



(**b**) Frame 441 of KITTI-CITY-71

Figure 6. Illustrations to some results in our numerical tests, with inliers in blue and outliers in red. (**a**) Frame 328 from KITTI-CITY-71, $N = 121766$; (**b**) Frame 441 from KITTI-CITY-71, $N = 119428$. Inliers/outliers are detected by using a ground-truth thresholding on the distance to the hyperplane recovered by each compared method. The results are represented by projecting 3D point clouds onto the image.

5. Conclusions

Sparse dictionary learning (SDL) and dual principal pursuit (DPCP) are two powerful tools in data science. In this paper, we formulate DPCP as a special case of the ℓ_m-norm maximization on the Stiefel manifold proposed for SDL. Then, we propose a novel smooth unconstrained penalty model PenNM for the original optimization problem with orthogonality constraints. We show PenNM is an exact penalty function of (1) under mild assumptions. We develop an novel approximate gradient approach PenNMF for solving PenNM. The global convergence of PenNMF as well as its sublinear convergence

rate are established. Numerical experiments illustrate that our proposed approach enjoys better performance than MSP/GPM [17,18] on various testing problems.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, project administration, X.H. and X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: Research is supported in part by the National Natural Science Foundation of China (No. 11971466, 11991021, and 11991020); Key Research Program of Frontier Sciences, Chinese Academy of Sciences (No. ZDBS-LY-7022); the National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences; and the Youth Innovation Promotion Association, Chinese Academy of Sciences.

Acknowledgments: We gratefully thank Yuexiang Zhai, Hermish Mehta, Zhengyuan Zhou, and Yi Ma for sharing their codes on MSP. Besides, we also gratefully thank Tianyu Ding, Zhihui Zhu, Tianjiao Ding, Yunchen Yang, Rene Vidal, Manolis Tsakiris, and Daniel Robinson for sharing their codes on DPCP problems.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DPCP	dual principal pursuit
DPCP-d	Denoised-DPCP
DPCP-IRLS	Iteratively-Reweighted-Least-Squares algorithm
DPCP-PSGM	Projected subgradient-based algorithm for solving DPCP
GPM	gradient projection method
MSP	matching, stretching, and projection
PenNM	penalty model for ℓ_m-norm maximization
PenNMF	first-order algorithm for solving our penalty model
RANSAC	Random Sampling and Consensus
SDL	Sparse dictionary learning

Appendix A. Proof for Theorem 2 and Corollary 3

In this section, we present the proof for Theorem 2 and Corollary 3. As (6) is an unconstrained optimization problem, the upper-bound for $\|\nabla h(W)\|_{\mathrm{F}}$ should be estimated. Before estimating the upper bounds of $\|\nabla f(W)\|_{\mathrm{F}}$ and $\left\|\nabla^2 f(W)[W(W^\top W - I_p)]\right\|_{\mathrm{F}}$, we first present two linear algebraic inequalities:

Lemma A1. *For any $A, B \in \mathbb{R}^{n\times N}$, $\|A \circ B\|_{\mathrm{F}} \le \|A\|_{\mathrm{F}} \|B\|_{\mathrm{F}}$.*

Proof.

$$\|A \circ B\|_{\mathrm{F}}^2 = \sum_i \sum_j A_{ij}^2 B_{ij}^2 \le \left(\sum_i \sum_j A_{ij}^2\right)\left(\sum_i \sum_j B_{ij}^2\right) = \|A\|_{\mathrm{F}}^2 \|B\|_{\mathrm{F}}^2 .$$

□

Lemma A2. *For any $A \in \mathbb{R}^{n\times N}$ and any $m \ge 3$, we have $\left\|A^{\circ(m)}\right\|_{\mathrm{F}} \le \|A\|_{\mathrm{F}}^m$.*

Proof. This lemma directly follows the fact that

$$\left\|A^{\circ(m)}\right\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{N} A_{i,j}^{2m} \le \left(\sum_{i=1}^{n}\sum_{j=1}^{N} A_{i,j}^2\right)^m = \|A\|_{\mathrm{F}}^{2m} .$$

□

Now, we present the upper bound estimation for $\|\nabla f(W)\|_{\mathrm{F}}$ and $\left\|\nabla^2 f(W)[W(W^\top W - I_p)]\right\|_{\mathrm{F}}$.

Lemma A3. *For any* $W \in \mathbb{R}^{n\times p}$, $\|\nabla f(W)\|_{\mathrm{F}} \le \|Y\|_{\mathrm{F}}^{m} \|W\|^{m-1}$.

Proof. Due to the fact that $\nabla f(W) = Y\left[(Y^\top W)^{\circ(m-1)} \circ \operatorname{sign}(Y^\top W)\right]$,

$$\begin{aligned}\|\nabla f(W)\|_{\mathrm{F}} &\le \|Y\|_{\mathrm{F}} \left\|(Y^\top W)^{\circ(m-1)} \circ \operatorname{sign}(Y^\top W)\right\|_{\mathrm{F}} \\ &= \|Y\|_{\mathrm{F}} \left\|(Y^\top W)^{\circ(m-1)}\right\|_{\mathrm{F}} \\ &\le \|Y\|_{\mathrm{F}} \left\|Y^\top W\right\|_{\mathrm{F}}^{m-1} \\ &\le \|Y\|_{\mathrm{F}}^{m} \|W\|^{m-1}.\end{aligned}$$

Here, the last inequality follows the fact that $\|AB\|_{\mathrm{F}} \le \|A\|_{\mathrm{F}} \|B\| \le \|A\|_{\mathrm{F}} \|B\|_{\mathrm{F}}$. □

Lemma A4. *For any* $W \in \mathbb{R}^{n\times p}$,

$$\left\|\nabla^2 f(W)[W(W^\top W - I_p)]\right\|_{\mathrm{F}} \le (m-1)\|W\|^{m-2} \left\|W(W^\top W - I_p)\right\| \|Y\|_{\mathrm{F}}^{m}.$$

Proof. From the expression of $\nabla^2 f(W)$ in Proposition 1,

$$\begin{aligned}&\left\|\nabla^2 f(W)[W(W^\top W)]\right\|_{\mathrm{F}} \\ &= (m-1)\left\|Y\left[(Y^\top W)^{\circ(m-2)} \circ (Y^\top W(W^\top W - I_p))\right]\right\|_{\mathrm{F}} \\ &\le (m-1)\|Y\|_{\mathrm{F}} \left\|(Y^\top W)^{\circ(m-2)} \circ (Y^\top W(W^\top W - I_p))\right\|_{\mathrm{F}} \\ &\le (m-1)\|Y\|_{\mathrm{F}} \left\|(Y^\top W)^{\circ(m-2)}\right\|_{\mathrm{F}} \left\|Y^\top W(W^\top W - I_p)\right\|_{\mathrm{F}} \\ &\le (m-1)\|W\|^{m-1} \left\|W(W^\top W - I_p)\right\| \|Y\|_{\mathrm{F}}^{m}.\end{aligned}$$

Here, the second inequality directly uses Lemma A1 and the last inequality follows Lemma A2. □

In the rest of this section, we consider the equivalence between (1) and (6). We first establish the relationships between the first-order stationary points of (6) and problem (1).

From the optimality condition of (6), we derive an important equality in Lemma A5.

Lemma A5. *For any first-order stationary point* $\tilde{W}$ *of* (6) *and any symmetric matrix* $T \in \mathbb{R}^{p\times p}$ *that satisfies* $T\tilde{W}^\top\tilde{W} = \tilde{W}^\top\tilde{W}T$, *we have*

$$\begin{aligned}0 =& \operatorname{tr}\left(T(\tilde{W}^\top\tilde{W} - I_p)\left(\beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) - \frac{3}{2}\Lambda(\tilde{W})\right)\right) \\ &- \frac{1}{2}\operatorname{tr}\left(T\tilde{W}^\top\nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)]\right).\end{aligned} \tag{A1}$$

Proof. Suppose $\tilde{W}$ is a first-order stationary point of (6), by the first-order optimality condition, $\nabla h(\tilde{W}) = 0$. Then, for any symmetric matrix $T \in \mathbb{R}^{p\times p}$ that satisfies $T\tilde{W}^\top\tilde{W} = \tilde{W}^\top\tilde{W}T$, $\langle \tilde{W}T, \nabla h(\tilde{W})\rangle = 0$.

As described in Proposition 1, $\nabla f(W)$ can be separated into three parts, we estimate their inner-product with $\tilde{W}T$ respectively.

First,

$$\begin{aligned}
&\left\langle \tilde{W}T, \nabla f(\tilde{W})\left(\frac{3}{2}I_p - \frac{1}{2}\tilde{W}^\top\tilde{W}\right) - \tilde{W}\Phi(\tilde{W}^\top\nabla f(\tilde{W}))\right\rangle \\
&= \mathrm{tr}\left(T\tilde{W}^\top\nabla f(\tilde{W})\left(\frac{3}{2}I_p - \frac{1}{2}\tilde{W}^\top\tilde{W}\right) - T\tilde{W}^\top\tilde{W}\Lambda(\tilde{W})\right) \\
&= \mathrm{tr}\left(T\left(\frac{3}{2}I_p - \frac{1}{2}\tilde{W}^\top\tilde{W}\right)\Lambda(\tilde{W}) - T\tilde{W}^\top\tilde{W}\Lambda(\tilde{W})\right) \\
&= \frac{3}{2}\mathrm{tr}\left(T(I_p - \tilde{W}^\top\tilde{W})\Lambda(\tilde{W})\right).
\end{aligned}$$

Here, the second equality follows the fact that that $\mathrm{tr}\,(BC) = \mathrm{tr}\,(BCB) = 0$ holds for any symmetric B and skew-symmetric C. Besides, the last inequality follows that $(\tilde{W}^\top\tilde{W} - I_p)T = T(\tilde{W}^\top\tilde{W} - I_p)$. As a result, we achieve the following equality:

$$\begin{aligned}
&\left\langle \tilde{W}T, \nabla f(\tilde{W})\left(\frac{3}{2}I_p - \frac{1}{2}\tilde{W}^\top\tilde{W}\right) - \tilde{W}\Phi(\tilde{W}^\top\nabla f(\tilde{W}))\right\rangle - \frac{1}{2}\left\langle \tilde{W}T, \nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)]\right\rangle \\
= &-\frac{3}{2}\mathrm{tr}\left(T(\tilde{W}^\top\tilde{W} - I_p)\Lambda(\tilde{W})\right) - \frac{1}{2}\left\langle T\tilde{W}, \nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)]\right\rangle.
\end{aligned}$$

Additionally, we estimate their inner-product of $\tilde{W}T$ and $\beta\tilde{W}(\tilde{W}^\top\tilde{W}\tilde{W}^\top\tilde{W} - I_p)$ and achieve the following equality

$$\begin{aligned}
&\left\langle \tilde{W}T, \beta(\tilde{W}^\top\tilde{W}\tilde{W}^\top\tilde{W} - I_p)\right\rangle \\
&= \left\langle \tilde{W}T, \beta\tilde{W}(\tilde{W}^\top\tilde{W} + I_p)(\tilde{W}^\top\tilde{W} - I_p)\right\rangle \\
&= \beta\mathrm{tr}\left(T\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p)(\tilde{W}^\top\tilde{W} - I_p)\right).
\end{aligned}$$

Based on the above two equations, multiplying $(\tilde{W}^\top\tilde{W} - I_p)\tilde{W}^\top$ on both sides of $0 = \nabla h(\tilde{W})$ results in

$$\begin{aligned}
0 &= \left\langle \tilde{W}T, \nabla h(\tilde{W})\right\rangle \\
&= \mathrm{tr}\left(T(\tilde{W}^\top\tilde{W} - I_p)\left(\beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) - \frac{3}{2}\Lambda(\tilde{W})\right)\right) \\
&\quad - \frac{1}{2}\mathrm{tr}\left(T\tilde{W}^\top\nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)]\right),
\end{aligned} \tag{A2}$$

and thus we complete the proof. □

Then based on the equality in Lemma A5, the following proposition shows that all first-order stationary point of (6) is uniformly bounded.

Proposition A6. *For any first-order stationary point $\tilde{W}$ of (6), suppose $\beta \geq (4m + 8)\|Y\|_{\mathrm{F}}^m$, then $\left\|\tilde{W}\right\|^2 \leq 1 + \frac{(m+2)\|Y\|_{\mathrm{F}}^m}{\beta}$.*

Proof. Let u denotes the top eigenvector of $\tilde{W}^\top\tilde{W}$, i.e. $\tilde{W}^\top\tilde{W}u = \left\|\tilde{W}^\top\tilde{W}\right\| u$.

Suppose $\tilde{W}$ is a first-order stationary point that satisfies $\left\|\tilde{W}\right\|^2 > 1 + \frac{(m+2)\|Y\|_{\mathrm{F}}^m}{\beta}$. By Lemma A5 we first have

$$
\begin{aligned}
0 &= \left\langle uu^\top \tilde{W}, \nabla f(\tilde{W}) \right\rangle \\
&= \operatorname{tr}\left(\beta uu^\top (\tilde{W}^\top \tilde{W} - I_p)\tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p)\right) - \frac{3}{2}\operatorname{tr}\left(uu^\top (\tilde{W}^\top \tilde{W} - I_p)\Lambda(\tilde{W})\right) \\
&\quad - \frac{1}{2}\operatorname{tr}\left(uu^\top \tilde{W}^\top \nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top \tilde{W} - I_p)]\right) \\
&\geq \beta\left(\left\|\tilde{W}\right\|^6 - \left\|\tilde{W}\right\|^2\right) - \frac{3}{2}\left\|\tilde{W}\right\|^2 u^\top \Lambda(\tilde{W})u - \frac{m-1}{2}\left\|Y\right\|_{\mathrm{F}}^m \left\|\tilde{W}\right\|^{m+2} \\
&\geq \beta\left(\left\|\tilde{W}\right\|^6 - \left\|\tilde{W}\right\|^2\right) - \frac{(m+2)\left\|Y\right\|_{\mathrm{F}}^m \left\|\tilde{W}\right\|^{m+2}}{2} \\
&\geq \left(\beta - \frac{(m+2)\left\|Y\right\|_{\mathrm{F}}^m}{2}\right)\left\|\tilde{W}\right\|^6 - \beta\left\|\tilde{W}\right\|^4 \\
&= \left\|\tilde{W}\right\|^4 \left[\left(\beta - \frac{(m+2)\left\|Y\right\|_{\mathrm{F}}^m}{2}\right)\left\|\tilde{W}\right\|^2 - \beta\right] \\
&> 0,
\end{aligned}
\tag{A3}
$$

which leads to the contradictory and shows that $\left\|\tilde{W}\right\|^2 \leq 1 + \frac{(m+2)\left\|Y\right\|_{\mathrm{F}}^m}{\beta}$. Here, the second equality directly follows Lemma A5. The first inequality uses Lemma A4 and the fact that $\left\|\tilde{W}(\tilde{W}^\top \tilde{W} - I_p)\right\| \leq \left\|\tilde{W}\right\|^3$. Besides, the second inequality follows the fact that $\tilde{W}^\top \tilde{W} \succeq \left\|\tilde{W}\right\|^2 uu^\top$. The second inequality uses the fact that $u^\top \Lambda(\tilde{W})u \leq \left\|\Lambda(\tilde{W})\right\|_{\mathrm{F}} \leq \left\|\tilde{W}^\top \nabla f(\tilde{W})\right\|_{\mathrm{F}} \leq \left\|\tilde{W}\right\|_2 \left\|\nabla f(\tilde{W})\right\|_{\mathrm{F}}$. The fourth inequality uses the fact that $\operatorname{tr}\left(\tilde{W}^\top \tilde{W} - I_p\right) \leq \left\|\tilde{W}\right\|_{\mathrm{F}}^2$, and the last inequality follows the fact that $\left\|\tilde{W}\right\|^2 - 1 \geq \frac{(m+2)\left\|Y\right\|_{\mathrm{F}}^m}{\beta}$. □

Combine Lemma A5 and Proposition A6, we restate Theorem 2 as Theorem A7 and achieve the equivalence between (1) and (6).

Theorem A7. *Suppose $\beta \geq (4m+8)\left\|Y\right\|_{\mathrm{F}}^m$, and $\tilde{W}$ is a first-order stationary point of (6), then either $\tilde{W}^\top \tilde{W} = I_p$ holds, which further implies that $\tilde{W}$ is a first-order stationary point of problem (1), or the inequality $\sigma_{\min}(\tilde{W}^\top \tilde{W}) \leq \frac{(2m+4)\left\|Y\right\|_{\mathrm{F}}^m}{\beta}$ holds.*

Proof. When $\beta \geq 4(m+2)\left\|Y\right\|_{\mathrm{F}}^m$, any first-order stationary point $\tilde{W}$ of (6) satisfies that $\left\|\tilde{W}\right\|^2 \leq 2$.

Suppose $\tilde{W}$ satisfies $\tilde{W}^\top \tilde{W} \succeq \frac{(2m+4)\left\|Y\right\|_{\mathrm{F}}^m}{\beta} I_p$, then $\frac{\beta}{2}\tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p) \succeq (m+2)\left\|\tilde{W}\right\|^m \left\|Y\right\|_{\mathrm{F}}^m \cdot I_p$. Then from Lemma A5, we have

$$
\begin{aligned}
0 &= \operatorname{tr}\left((\tilde{W}^\top \tilde{W} - I_p)^2\left(\beta\tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p) - \frac{3}{2}\Lambda(\tilde{W})\right)\right) \\
&\quad - \frac{1}{2}\operatorname{tr}\left((\tilde{W}^\top \tilde{W} - I_p)\tilde{W}^\top \nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top \tilde{W} - I_p)]\right) \\
&\geq \operatorname{tr}\left((\tilde{W}^\top \tilde{W} - I_p)^2\left(\beta\tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p) - \frac{m+2}{2}\left\|W\right\|^m \left\|Y\right\|_{\mathrm{F}}^m \cdot I_p\right)\right) \\
&\geq \operatorname{tr}\left(\beta(\tilde{W}^\top \tilde{W} - I_p)^2 \tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p)\right) \\
&\geq 0,
\end{aligned}
$$

showing that $\operatorname{tr}\left(\beta(\tilde{W}^\top \tilde{W} - I_p)^2 \tilde{W}^\top \tilde{W}(\tilde{W}^\top \tilde{W} + I_p)\right) = 0$. Then by the positive-definiteness of $\tilde{W}^\top \tilde{W}$, we can conclude that $\tilde{W}^\top \tilde{W} - I_p = 0$.

As a result, we have that either $\tilde{W}^\top \tilde{W} - I_p = 0$ or $\sigma_{\min}(\tilde{W}^\top \tilde{W}) \leq \frac{(2m+4)\left\|Y\right\|_{\mathrm{F}}^m}{\beta}$, and completes the proof. □

Corollary A8. *Suppose $p = 1$ in (1), $\beta \geq (4m+8)\|Y\|_F^m$, and $\tilde{W}$ is a first-order stationary point of (6), then either $\tilde{W}^\top \tilde{W} = I_p$ holds, which further implies that $\tilde{W}$ is a first-order stationary point of problem (1), or $\tilde{W} = 0$.*

Proof. By the same routine of Theorem 2, when $\beta \geq 4(m+2)\|Y\|_F^m$, any first-order stationary point $\tilde{W}$ of (6) satisfies that $\left\|\tilde{W}\right\|^2 \leq 2$. Then, following the same proof routine in Lemma A5 and Theorem 2, we have

$$\begin{aligned} 0 &\geq \operatorname{tr}\left((\tilde{W}^\top\tilde{W} - I_p)^2\left(\beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) - \frac{m+2}{2}\|W\|^m\|Y\|_F^m \cdot I_p\right)\right) \\ &= \left(\left\|\tilde{W}\right\|^2 - 1\right)^2\left(\beta\left\|\tilde{W}\right\|^2\left(\left\|\tilde{W}\right\|^2 + I_p\right) - \frac{m+2}{2}\|W\|^m\|Y\|_F^m\right) \\ &\geq 0. \end{aligned}$$

When $\beta \geq \frac{4m+8}{2}\|Y\|_F$, we have

$$\frac{\beta}{2}\|W\|_F^4 + \frac{\beta}{2}\|W\|_F^2 > \frac{m+2}{2}\|Y\|_F\left\|\tilde{W}\right\|_2^m \tag{A4}$$

holds for any $W \in \mathbb{R}^n \setminus 0$.

Then, for any $\tilde{W} \neq 0$, we can conclude that

$$\left(\beta\left\|\tilde{W}\right\|^2\left(\left\|\tilde{W}\right\|^2 + I_p\right) - \frac{m+2}{2}\|W\|^m\|Y\|_F^m\right) > 0,$$

and thus $\left\|\tilde{W}\right\|^2 - 1 = 0$. As a result, from (A1), when $\tilde{W}$ is a first-order stationary point of (6), either $\tilde{W} = 0$ or $\tilde{W}^\top\tilde{W} - I_p = 0$. □

Appendix B. Proof for Theorem 4

In this section, we present the main body of the proof for Theorem 4. To show the convergence of PenNMF, we first present some preliminary lemmas. Then, we show that the updating direction $D(W^k)$ is a descending direction and thus $h(W^{k+1}) \leq h(W^k)$, as illustrated in Lemma A12. Together with Lemma A10, we show that the sequence is restricted in the neighborhood of the constraints, and we achieve the global convergence property of PenNMF in Theorem 4. We first estimate the upper-bound of the term $\left|f(W) - \frac{1}{2}\left\langle W^\top W - I_p, \Lambda(W)\right\rangle\right|$ in $h(W)$.

Lemma A9. *For any $W \in \mathbb{R}^{n\times p}$,*

$$\left|f(W) - \frac{1}{2}\left\langle W^\top W - I_p, \Lambda(W)\right\rangle\right| \leq \frac{m}{2}\|Y\|_F^m \max\{\|W\|^2 + 1, 2\}\|W\|^m.$$

Proof. We first estimate the upper-bound for $|f(W)|$, which can be achieved by

$$|f(W)| = \frac{1}{m}\left\|W^\top Y\right\|_m^m \leq \frac{1}{m}\left\|W^\top Y\right\|_F^m \leq \frac{1}{m}\|Y\|_F^m\|W\|^m.$$

Besides, from Lemma A3, we have

$$\begin{aligned}
&\left|\frac{1}{2}\left\langle W^\top W - I_p, \Lambda(W)\right\rangle\right| = \frac{1}{2}\left|\left\langle W(W^\top W - I_p), \nabla f(W)\right\rangle\right| \\
\leq\ & \frac{1}{2}\left\|W(W^\top W - I_p)\right\|_{\mathrm{F}} \|\nabla f(W)\|_{\mathrm{F}} \\
\leq\ & \frac{m}{2}\left\|W(W^\top W - I_p)\right\|_2 \|Y\|_{\mathrm{F}}^m \|W\|^{m-1} \\
\leq\ & \frac{m}{2}\|Y\|_{\mathrm{F}}^m \max\{1, \|W\|^2\} \|W\|^m .
\end{aligned}$$

Combine the above two equations, we achieve

$$|f(W) - \frac{1}{2}\left\langle W^\top W - I_p, \Lambda(W)\right\rangle| \leq \frac{m}{2}\|Y\|_{\mathrm{F}}^m \max\{\|W\|^2 + 1, 2\} \|W\|^m ,$$

and complete the proof. □

We then show that the penalty term $\psi(W) := \frac{1}{6}\|W\|_{\mathrm{F}}^6 - \frac{1}{2}\|W\|_{\mathrm{F}}^2$ builds a barrier around $\mathcal{S}_{n,p}$, i.e., those points that are sufficiently far from $\mathcal{S}_{n,p}$ have higher functional value than those points that are close to $\mathcal{S}_{n,p}$.

Lemma A10. *Suppose for any* $\delta \in \left(0, \frac{1}{3}\right]$ *and* $\beta \geq \max\{228m\|Y\|_{\mathrm{F}}^m, \frac{32m}{\delta}\|Y\|_{\mathrm{F}}^m\}$, *we have*

$$\max_{\|W^\top W - I_p\|_{\mathrm{F}}^2 \leq \frac{\delta}{8}} h(W) \leq \min_{\|W^\top W - I_p\|_{\mathrm{F}}^2 \geq \delta} h(W). \tag{A5}$$

Proof. Let $\psi(W) := \frac{1}{6}\left(\|W\|^6 - \|W\|^2\right)$. For any W_1 satisfies $\psi(W_1) \leq \delta$ and W_2 satisfies $\|W_2\|^2 \leq 2$ and $\psi(W_2) \geq 2\delta$, then

$$\begin{aligned}
&h(W_2) - h(W_1) \\
\geq\ & \beta\delta - |\nabla f(W_1)| - \frac{1}{2}\left\langle W_1^\top W_1 - I_p, \Lambda(W_1)\right\rangle| - |\nabla f(W_2)| - \frac{1}{2}\left\langle W_2^\top W_2 - I_p, \Lambda(W_2)\right\rangle| \\
\geq\ & \beta\delta - m\|Y\|_{\mathrm{F}}^m (1+\delta)^{m+2} - m\|Y\|_{\mathrm{F}}^m 2^{m+2} \\
\geq\ & \beta\delta - 19m\|Y\|_{\mathrm{F}}^m \\
\geq\ & 0.
\end{aligned} \tag{A6}$$

Here, the second inequality uses the fact that $\|W_1\|_2^2 \leq 1 + \left\|W_1^\top W_1 - I_p\right\|_{\mathrm{F}} \leq 1 + \delta$, and $\|W_2\|^2 \leq 2$. Moreover, when $\|W_2\|_{\mathrm{F}}^2 \geq 2$, we have $\psi(W_2) \geq \frac{\beta}{6}\left\|\tilde{W}\right\|^6$. Then, $\frac{1}{2}\psi(W_2) \geq \frac{1}{3}\beta \geq \delta\beta \geq \psi(W_1)$.

$$\begin{aligned}
h(W_2) - h(W_1) \geq\ & -m\|Y\|_{\mathrm{F}}^m (1+\delta)^{m+2} + \psi(W_2) - 16m\|Y\|_{\mathrm{F}}^m \|W\|^{m+2} - \delta\beta \\
\geq\ & -19m\|Y\|_{\mathrm{F}}^m \|W_2\|^{m+2} + \frac{\beta}{12}\left\|\tilde{W}\right\|^6 \\
\geq\ & 0.
\end{aligned} \tag{A7}$$

Here, the second inequality follows the fact that $(1+\delta)^m \leq (1+\delta)^4 < 3$.
Besides, as $\left\|W^\top W - I_p\right\|_{\mathrm{F}}^2 \leq \delta$ implies

$$\frac{1}{6}\|W\|_{\mathrm{F}}^6 - \frac{1}{2}\|W\|_{\mathrm{F}}^2 + \frac{p}{3} = \frac{1}{6}\mathrm{tr}\left[(W^\top W - I_p)^2(W^\top W + 2I_p)\right]. \tag{A8}$$

As a result, $\psi(W) \leq 3\delta$ implies $\left\|W^\top W - I_p\right\|_{\mathrm{F}} \leq \delta$. Besides, $\left\|W^\top W - I_p\right\|_{\mathrm{F}}^2 \leq \delta$ implies $\psi(W) \leq \frac{2}{3}\delta$. Therefore,

$$\max_{\|W^\top W - I_p\|_{\mathrm{F}}^2 \leq \frac{\delta}{8}} h(W) \leq \min_{\|W^\top W - I_p\|_{\mathrm{F}}^2 \geq \delta} h(W). \tag{A9}$$

□

Lemma A10 shows that the smooth penalty term builds a barrier around $\mathcal{S}_{n,p}$. Moreover, we characterize the relations between $\|D(W)\|_{\mathrm{F}}$ and $\left\|W^\top W - I_p\right\|_{\mathrm{F}}$ in the following lemma.

Lemma A11. *Suppose* $\delta \in \left(0, \frac{1}{3}\right]$, *set* $||W^\top W - I_p||_{\mathrm{F}}^2 \leq \frac{1}{8}\delta$, *and* $\beta \geq (4m+8)\|Y\|_{\mathrm{F}}^m$. *Then,*

$$\|D(W)\|_{\mathrm{F}} \geq \frac{\sqrt{3}\beta}{6} \cdot \left\|W^\top W - I_p\right\|_{\mathrm{F}}, \tag{A10}$$

where $D(W) := \nabla f(W) - W\Phi(W^\top \nabla f(W)) + \beta W(W^\top W W^\top W - I_p)$.

Proof. First, we present two linear algebra relationships: The first is the inequality $||A||_{\mathrm{F}} \geq \left\|\frac{A+A^\top}{2}\right\|_{\mathrm{F}}$ holds for any square matrix A, which is quite obvious and the proof is omitted. The second is the equality $||AB + BA||_{\mathrm{F}} = 2||AB||_{\mathrm{F}}$ holds for any symmetric matrices A and B, which results from the fact $||AB + BA||_{\mathrm{F}}^2 = 2||AB||_{\mathrm{F}}^2 + 2\mathrm{tr}(ABAB) = 2||AB||_{\mathrm{F}}^2 + 2\mathrm{tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}}A^{\frac{1}{2}}BA^{\frac{1}{2}}) = 4||AB||_{\mathrm{F}}^2$.

It follows from the above facts that

$$\begin{aligned}
&\left\|W^\top D(W)\right\|_{\mathrm{F}} \geq \frac{1}{2}\left\|W^\top D(W) + D(W)^\top W\right\|_{\mathrm{F}} \\
=&\frac{1}{2}\left\|\left(\beta W^\top W(W^\top W + I_p) - \Lambda(W)\right)(W^\top W - I_p) + (W^\top W - I_p)\left(\beta W^\top W(W^\top W + I_p) - \Lambda(W)\right)\right\|_{\mathrm{F}} \\
\geq&\frac{\beta}{3} \cdot ||W^\top W - I_p||_{\mathrm{F}},
\end{aligned}$$

where the last equality uses the fact that $\sigma_{\min}(W^\top W(W^\top W + I_p)) \geq \sigma_{\min}(W^\top W) \geq 1 - \delta \geq \frac{2}{3}$.

Together with the facts that $\left\|W^\top D(W)\right\|_{\mathrm{F}} \leq \|W\|_2 \|D(W)\|_{\mathrm{F}}$ and $\sigma_{\max}(W^\top W) \leq 1 + \delta \leq \frac{4}{3}$, we have

$$\begin{aligned}
\|D(W)\|_{\mathrm{F}} \geq&\frac{1}{\|W\|_{\mathrm{F}}}\left\|W^\top D(W)\right\|_{\mathrm{F}} \geq \frac{\sqrt{3}}{2}\left\|W^\top D(W)\right\|_{\mathrm{F}} \\
\geq&\frac{\sqrt{3}\beta}{6}\left\|W^\top W - I_p\right\|_{\mathrm{F}}.
\end{aligned}$$

□

Let $M_1 := \sup \frac{\|\nabla h(W_1) - \nabla h(W_2)\|_{\mathrm{F}}}{\|W_1 - W_2\|_{\mathrm{F}}}$, then we have that the following illustrating that PenNMF generates a descending sequence $\{W_k\}$.

Lemma A12. *Suppose* $\delta \in \left(0, \frac{1}{3}\right]$ *and* $\beta \geq \max\left\{228m\|Y\|_{\mathrm{F}}^m, \frac{32}{\delta}\|Y\|_{\mathrm{F}}^m\right\}$. *Let* $\{W^k\}$ *be the iterate sequence generated by PenNMF, starting from any initial point* W^0 *satisfying* $||{W^0}^\top W^0 - I_p||_{\mathrm{F}}^2 \leq \frac{1}{8}\delta$, *and the stepsize* $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, *where* $\bar{\eta} \leq \frac{1}{2M_1}$. *Then, it holds that*

$$h(W^{k+1}) \leq h(W^k) - \frac{\bar{\eta}}{4}\left\|D(W^k)\right\|_{\mathrm{F}}^2 \tag{A11}$$

for any $k = 0, 1, \cdots$.

Proof. By the explicit expression of $\nabla h(W^k)$, we first have

$$\begin{aligned}
&\left\|\nabla h(W^k) - D(W^k)\right\|_{\mathrm{F}} \\
=& \left\|\frac{1}{2}\nabla f(W^k)({W^k}^\top W^k - I_p) + \frac{1}{2}\nabla^2 f(W^k)[W^k({W^k}^\top W^k - I_p)]\right\|_{\mathrm{F}} \\
\leq& \frac{1}{2}\left[\left\|\nabla f(W^k)({W^k}^\top W^k - I_p)\right\|_{\mathrm{F}} + \left\|\nabla^2 f(W^k)[W^k({W^k}^\top W^k - I_p)]\right\|_{\mathrm{F}}\right] \\
\leq& \frac{m}{2}\left\|Y\right\|_{\mathrm{F}}^m \left\|W^k\right\|^{m-1} \left\|{W^k}^\top W^k - I_p\right\|_{\mathrm{F}} \\
\leq& 2m\left\|Y\right\|_{\mathrm{F}}^m \left\|{W^k}^\top W^k - I_p\right\|_{\mathrm{F}}.
\end{aligned} \tag{A12}$$

Here, the first inequality follows Lemma A3 and Lemma A4.
Besides, by the definition of M_1, we have

$$h(W^{k+1}) \leq h(W^k) + \left\langle W^{k+1} - W^k, \nabla h(W^k)\right\rangle + \frac{M_1}{2}\left\|W^{k+1} - W^k\right\|_{\mathrm{F}}^2. \tag{A13}$$

Suppose $\left\|{W^k}^\top W^k - I_p\right\|_{\mathrm{F}}^2 \leq \delta$, then by Lemma A11 we can conclude that

$$\left\|\nabla h(W^k) - D(W^k)\right\|_{\mathrm{F}} \leq 2m\left\|Y\right\|_{\mathrm{F}}^m \left\|{W^k}^\top W^k - I_p\right\|_{\mathrm{F}} \leq \frac{4\sqrt{3}m\left\|Y\right\|_{\mathrm{F}}^m}{\beta}\left\|D(W^k)\right\|_{\mathrm{F}}. \tag{A14}$$

Substitute $W^{k+1} - W^k = -\eta D(W^k)$ and (A14) into (A13), we have

$$\begin{aligned}
&h(W^{k+1}) - h(W^k) \\
\leq& \left\langle W^{k+1} - W^k, \nabla h(W^k)\right\rangle + \frac{M_1}{2}\left\|W^{k+1} - W^k\right\|_{\mathrm{F}}^2 \\
\leq& -\eta^k\left\langle D(W^k), D(W^k)\right\rangle + \frac{M_1}{2}\left\|\eta^k D(W^k)\right\|_{\mathrm{F}}^2 + \left|\left\langle \nabla h(W^k) - D(W^k), D(W^k)\right\rangle\right| \\
\leq& \left(-\eta^k + \frac{4\sqrt{3}m\left\|Y\right\|_{\mathrm{F}}^m}{\beta}\eta^k + \frac{M_1}{2}\left(\eta^k\right)^2\right)\left\|D(W^k)\right\|_{\mathrm{F}}^2 \\
\leq& -\frac{\eta^k}{2}\left\|D(W^k)\right\|_{\mathrm{F}}^2 \leq -\frac{\bar{\eta}}{4}\left\|D(W^k)\right\|_{\mathrm{F}}^2.
\end{aligned} \tag{A15}$$

Then by Lemma A10, as $h(W^{k+1}) \leq h(W^k)$, we can conclude that $\left\|{W^{k+1}}^\top W^{k+1} - I_p\right\|_{\mathrm{F}}^2 \leq \delta$. Then, by induction we can conclude that $\left\|{W^k}^\top W^k - I_p\right\|_{\mathrm{F}}^2 \leq \delta$ holds for $k = 1, 2, 3, \cdots$. Then, by (A15) again we conclude that

$$h(W^{k+1}) \leq h(W^k) - \frac{\bar{\eta}}{4}\left\|D(W^k)\right\|_{\mathrm{F}}^2 \tag{A16}$$

holds for $k = 1, 2, 3, \cdots$ and completes our proof. □

The following lemma shows that when our algorithm stops at $\tilde{W}$, then $\tilde{W}$ is a first-order stationary point of (1).

Lemma A13. *Suppose* $\delta \in \left(0, \frac{1}{3}\right]$ *and* $\beta \geq \max\left\{228m\left\|Y\right\|_{\mathrm{F}}^m, \frac{32}{\delta}\left\|Y\right\|_{\mathrm{F}}^m\right\}$. *For any* $\tilde{W}$ *satisfying* $\left\|\tilde{W}^\top\tilde{W} - I_p\right\|_{\mathrm{F}}^2 \leq \delta$ *and* $D(\tilde{W}) = 0$, *we have that* $\tilde{W}$ *is a first-order stationary point of* (1).

Proof. Suppose $D(\tilde{W}) = 0$, then by the same proof routine in Theorem 2, we consider the inner-product of $D(\tilde{W})$ and $\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)$:

$$\begin{aligned}
0 &= \left\langle D(\tilde{W}), \tilde{W}(\tilde{W}^\top\tilde{W} - I_p)\right\rangle \\
&= \left\langle \nabla f(\tilde{W}) - \tilde{W}\Lambda(\tilde{W}) + \beta\tilde{W}(\tilde{W}^\top\tilde{W}\tilde{W}^\top\tilde{W} - I_p), \tilde{W}(\tilde{W}^\top\tilde{W} - I_p)\right\rangle \\
&= \mathrm{tr}\left((\tilde{W}^\top\tilde{W} - I_p)\tilde{W}^\top\nabla f(\tilde{W}) - (\tilde{W}^\top\tilde{W} - I_p)\tilde{W}^\top\tilde{W}\Lambda(\tilde{W}) + \beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p)(\tilde{W}^\top\tilde{W} - I_p)^2\right) \\
&= \mathrm{tr}\left((\tilde{W}^\top\tilde{W} - I_p)^2(\beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) - \Lambda(\tilde{W})\right) \\
&\geq \mathrm{tr}\left((\tilde{W}^\top\tilde{W} - I_p)^2(\beta\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) - (4m+8)\left\|Y\right\|_{\mathrm{F}}^m \cdot I_p\right) \\
&\geq \frac{\beta}{2}\mathrm{tr}\left((\tilde{W}^\top\tilde{W} - I_p)^2\right) \geq 0.
\end{aligned}$$

Here, the fourth equation follows the definition of $\Lambda(\tilde{W}) := \Phi(\tilde{W}^\top\nabla f(\tilde{W}))$ and the first inequality uses the fact that $\left\|\tilde{W}\right\| \leq 2$, then together with Lemma A3, we can conclude that $\left\|\Lambda(\tilde{W})\right\|_2 \preceq (4m+8)\left\|Y\right\|_{\mathrm{F}}^m \cdot I_p$. Besides, the last inequality uses the fact that $\frac{\beta}{2}\tilde{W}^\top\tilde{W}(\tilde{W}^\top\tilde{W} + I_p) \succeq (4m+8)\left\|Y\right\|_{\mathrm{F}}^m \cdot I_p$.

Then we can conclude that that $\tilde{W}^\top\tilde{W} = I_p$. By the definition of $\nabla h(\tilde{W})$, we have

$$\nabla h(\tilde{W}) - D(\tilde{W}) = -\frac{1}{2}\nabla f(\tilde{W})(\tilde{W}^\top\tilde{W} - I_p) - \frac{1}{2}\nabla^2 f(\tilde{W})[\tilde{W}(\tilde{W}^\top\tilde{W} - I_p)] = 0, \tag{A17}$$

showing that $\nabla f(\tilde{W}) = 0$. Together with Theorem 2 we can conclude that $\tilde{W}$ is a first-order stationary point of (1). □

Based on the Lemmas A10–A13, we restate Theorem 4 as Theorem A14 and show the global convergence property of PenNMF in the following theorem.

Theorem A14. *Suppose* $\delta \in \left(0, \frac{1}{3}\right]$ *and* $\beta \geq \max\left\{228m\left\|Y\right\|_{\mathrm{F}}^m, \frac{32}{\delta}\left\|Y\right\|_{\mathrm{F}}^m\right\}$. *Let* $\{W^k\}$ *be the iterate sequence generated by PenNMF, starting from any initial point* W^0 *satisfying* $||{W^0}^\top W^0 - I_p||_{\mathrm{F}}^2 \leq \frac{1}{8}\delta$, *and the stepsize* $\eta_k \in \left[\frac{1}{2}\bar{\eta}, \bar{\eta}\right]$, *where* $\bar{\eta} \leq \frac{1}{2M_1}$. *Then,* W^k *weakly converges to a first-order stationary point of* (1). *Moreover, for any* $k = 1, 2, \cdots$, *the convergence rate of PenNMF can be estimated by*

$$\min_{0\leq i\leq k}\left\|D(W^i)\right\|_{\mathrm{F}} \leq \sqrt{\frac{8m\left\|Y\right\|_{\mathrm{F}}^m + 2\beta\delta}{\bar{\eta}(k+1)}}. \tag{A18}$$

Proof. By Lemma A12, it holds that

$$h(W^{k+1}) \leq h(W^k) - \frac{\bar{\eta}}{4}\left\|D(W^k)\right\|_{\mathrm{F}}^2.$$

If W^* is a cluster point of $\{W^k\}$, we have $W^{k+1} - W^k = 0$. Together with ${W^*}^\top W^* = I_p$ implied by Lemma A11, we can conclude that W^* is a first-order stationary point of problem (1).

Calculating the summation of the above inequalities from $k = 0$ to $N-1$, we have

$$\begin{aligned}
&\sum_{i=0}^{k}\frac{\bar{\eta}}{4}\left\|D(W^i)\right\|_{\mathrm{F}}^2 \leq h(W^0) - h(W^k) < h(W^0) - \inf_{||W^\top W - I_p||_{\mathrm{F}}^2 \leq \delta} h(W) \\
<& \sup_{||W^\top W - I_p||_{\mathrm{F}}^2 \leq \delta}\tilde{h}(W) - \inf_{||W^\top W - I_p||_{\mathrm{F}}^2 \leq \delta}\tilde{h}(W) + \frac{\beta}{4}\left(||{W^0}^\top W^0 - I_p||_{\mathrm{F}}^2 - ||{W^k}^\top W^k - I_p||_{\mathrm{F}}^2\right) \\
\leq& 2m\left\|Y\right\|_{\mathrm{F}}^m + \frac{\beta\delta}{2},
\end{aligned} \tag{A19}$$

showing that $\liminf_{k\to+\infty} \left\|D(W^k)\right\|_{\mathrm{F}} = 0$, which further implies that $D(W^*) = 0$. By Lemma A13, $D(W^*) = 0$ implies that W^* is a first-order stationary point of (1).

Moreover, by (A19), we have that

$$\min_{0\le i\le k} \left\|D(W^i)\right\|_{\mathrm{F}}^2 \le \frac{1}{k+1}\sum_{i=0}^{k} \left\|D(W^i)\right\|_{\mathrm{F}}^2 \le \frac{8m\,\|Y\|_{\mathrm{F}}^m + 2\beta\delta}{\bar{\eta}(k+1)},$$

and complete the proof. □

Appendix C. Additional Experimental Results

In this section, we propose some additional numerical experiments. Figure A1 shows the top 12 basis computed from the testing instances in Section 4.1 by PenNMF. As described in [17], the top bases are those with the largest coefficients in terms of ℓ_1-norm.

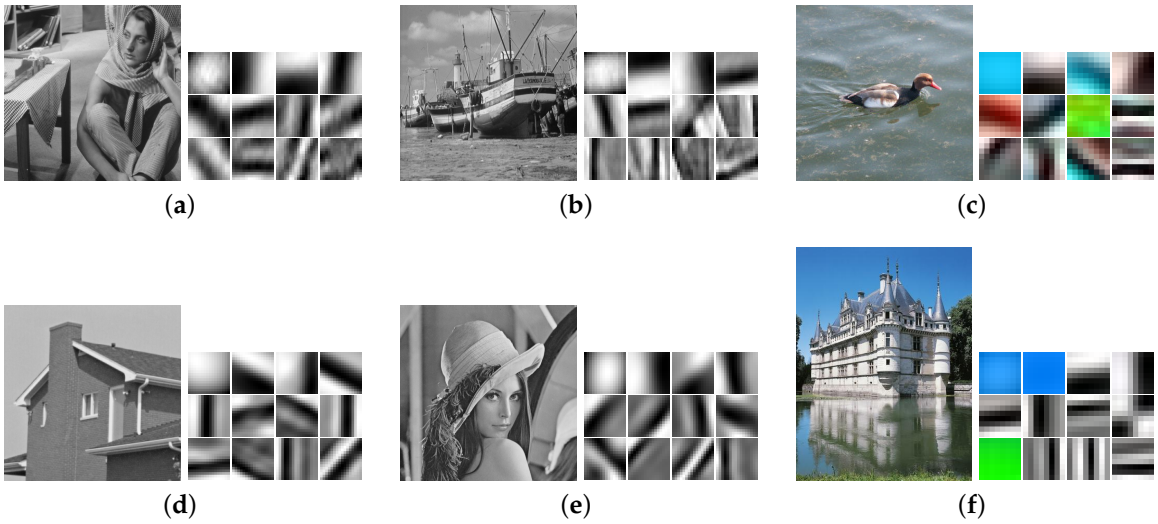


Figure A1. The top 12 bases computed from all patches of the test images without noise by PenNMF. (**a**) "Barbara"; (**b**) "Boat"; (**c**) "Duck"; (**d**) "House"; (**e**) "Lena"; (**f**) "Chateau".

References

1. Hansen, T.L.; Badiu, M.A.; Fleury, B.H.; Rao, B.D. A sparse Bayesian learning algorithm with dictionary parameter estimation. In Proceedings of the Sensor Array and Multichannel Signal Processing Workshop (SAM), A Coruña, Spain, 22–25 June 2014; pp. 385–388.
2. Shen, H.; Li, X.; Zhang, L.; Tao, D.; Zeng, C. Compressed Sensing-Based Inpainting of Aqua Moderate Resolution Imaging Spectroradiometer Band 6 Using Adaptive Spectrum-Weighted Sparse Bayesian Dictionary Learning. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 894–906. [CrossRef]
3. Bai, Y.; Jiang, Q.; Sun, J. Subgradient descent learns orthogonal dictionaries. *arXiv* **2018**, arXiv:1810.10702.
4. Gilboa, D.; Buchanan, S.; Wright, J. Efficient dictionary learning with gradient descent. *arXiv* **2018**, arXiv:1809.10313.
5. Kuo, H.W.; Zhang, Y.; Lau, Y.; Wright, J. Geometry and symmetry in short-and-sparse deconvolution. *SIAM J. Math. Data Sci.* **2020**, *2*, 216–245. [CrossRef]
6. Rambhatla, S.; Li, X.; Haupt, J. NOODL: Provable Online Dictionary Learning and Sparse Coding. *arXiv* **2019**, arXiv:1902.11261.
7. Song, X.; Wu, L. A Novel Hyperspectral Endmember Extraction Algorithm Based on Online Robust Dictionary Learning. *Remote Sens.* **2019**, *11*, 1792. [CrossRef]
8. Sun, J.; Qu, Q.; Wright, J. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Trans. Inf. Theory* **2016**, *63*, 853–884. [CrossRef]

9. Wang, D.; Wan, J.; Chen, J.; Zhang, Q. An Online Dictionary Learning-Based Compressive Data Gathering Algorithm in Wireless Sensor Networks. *Sensors* **2016**, *16*, 1547. [CrossRef]
10. Yang, L.; Fang, J.; Cheng, H.; Li, H. Sparse Bayesian dictionary learning with a Gaussian hierarchical model. *Signal Process.* **2017**, *130*, 93–104. [CrossRef]
11. Wang, Y.; Wu, S.; Yu, B. Unique Sharp Local Minimum in ℓ_1-minimization Complete Dictionary Learning. *arXiv* **2019**, arXiv:1902.08380.
12. Zhang, Y.; Kuo, H.W.; Wright, J. Structured local optima in sparse blind deconvolution. *IEEE Trans. Inf. Theory* **2019**, *66*, 419–452. [CrossRef]
13. Zhou, Q.; Feng, Z.; Benetos, E. Adaptive Noise Reduction for Sound Event Detection Using Subband-Weighted NMF. *Sensors* **2019**, *19*, 3206. [CrossRef] [PubMed]
14. Ling, Y.; Gao, H.; Zhou, S.; Yang, L.; Ren, F. Robust Sparse Bayesian Learning-Based Off-Grid DOA Estimation Method for Vehicle Localization. *Sensors* **2020**, *20*, 302. [CrossRef] [PubMed]
15. Liu, S.; Huang, Y.; Wu, H.; Tan, C.; Jia, J. Efficient Multi-Task Structure-Aware Sparse Bayesian Learning for Frequency-Difference Electrical Impedance Tomography. *IEEE Trans. Industr. Inform.* **2020**. [CrossRef]
16. Qu, Q.; Zhu, Z.; Li, X.; Tsakiris, M.C.; Wright, J.; Vidal, R. Finding the Sparsest Vectors in a Subspace: Theory, Algorithms, and Applications. *arXiv* **2020**, arXiv:2001.06970.
17. Zhai, Y.; Yang, Z.; Liao, Z.; Wright, J.; Ma, Y. Complete Dictionary Learning via ℓ_4-Norm Maximization over the Orthogonal Group. *arXiv* **2019**, arXiv:1906.02435.
18. Shen, Y.; Xue, Y.; Zhang, J.; Letaief, K.B.; Lau, V. Complete Dictionary Learning via ℓ_p-norm Maximization. *arXiv* **2020**, arXiv:2002.10043.
19. Gao, B.; Liu, X.; Yuan, Y.x. Parallelizable Algorithms for Optimization Problems with Orthogonality Constraints. *SIAM J. Sci. Comput.* **2019**, *41*, A1949–A1983. [CrossRef]
20. Wen, Z.; Yang, C.; Liu, X.; Zhang, Y. Trace-penalty minimization for large-scale eigenspace computation. *J. Sci. Comput.* **2016**, *66*, 1175–1203. [CrossRef]
21. Xiao, N.; Liu, X.; Yuan, X. A Class of Smooth Exact Penalty Function Methods for Optimization Problems with Orthogonality Constraints. Avaiable online: http://www.optimization-online.org/DB_HTML/2020/02/7607.html (accessed on 26 May 2020).
22. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Rob. Res.* **2013**, *32*, 1231–1237. [CrossRef]
23. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
24. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.
25. Xu, H.; Caramanis, C.; Sanghavi, S. Robust PCA via outlier pursuit. *IEEE Trans. Inf. Theory* **2010**, *58*, 3047–3064. [CrossRef]
26. Soltanolkotabi, M.; Candes, E.J. A geometric analysis of subspace clustering with outliers. *Ann. Stat.* **2012**, *40*, 2195–2238. [CrossRef]
27. Rahmani, M.; Atia, G.K. Coherence pursuit: Fast, simple, and robust principal component analysis. *IEEE Trans. Signal Process.* **2017**, *65*, 6260–6275. [CrossRef]
28. You, C.; Robinson, D.P.; Vidal, R. Provable self-representation based outlier detection in a union of subspaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3395–3404.
29. Ding, T.; Zhu, Z.; Ding, T.; Yang, Y.; Robinson, D.; Vidal, R.; Tsakiris, M. Noisy dual principal component pursuit. In Proceedings of the International Conference on Machine learning, Long Beach, CA, USA, 10–15 June 2019.
30. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
31. Tsakiris, M.C.; Vidal, R. Dual principal component pursuit. *J. Mach. Learn. Res.* **2018**, *19*, 684–732.
32. Zhu, Z.; Wang, Y.; Robinson, D.P.; Naiman, D.Q.; Vidal, R.; Tsakiris, M.C. Dual principal component pursuit: probability analysis and efficient algorithms. *arXiv* **2018**, arXiv:1812.09924.
33. Shi, L.; Chi, Y. Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *arXiv* **2019**, arXiv:1911.11167.

34. Qu, Q.; Li, X.; Zhu, Z. A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, CB, Canada, 2019; pp. 4017–4028.
35. Qu, Q.; Sun, J.; Wright, J. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QB, Canada, 8–13 December 2014; pp. 3401–3409.
36. Barzilai, J.; Borwein, J.M. Two-point step size gradient methods. *IMA J. Numer. Anal.* **1988**, *8*, 141–148. [CrossRef]
37. Dai, Y.H.; Fletcher, R. Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. *Numer. Math.* **2005**, *100*, 21–47. [CrossRef]
38. Absil, P.A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, USA, 2009.
39. Boumal, N.; Mishra, B.; Absil, P.A.; Sepulchre, R. Manopt, a Matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.* **2014**, *15*, 1455–1459.
40. Mairal, J.; Elad, M.; Sapiro, G. Sparse Representation for Color Image Restoration. *IEEE Trans. Image Process.* **2008**, *17*, 53–69.[CrossRef] [PubMed]
41. Dolan, E.D.; Moré, J.J. Benchmarking optimization software with performance profiles. *Math. Program.* **2002**, *91*, 201–213. [CrossRef]