Pushing the Limits of Affine Rank Minimization by Adapting Probabilistic PCA

Bo Xin BOXIN@PKU.EDU.CN
Nat'l Engineering Laboratory for Video Technology; Cooperative Medianet Innovation Center, Peking University, Beijing

David Wipf DAVIDWIP@MICROSOFT.COM
Microsoft Research, Beijing

Abstract

Many applications require recovering a matrix of minimal rank within an affine constraint set, with matrix completion a notable special case. Because the problem is NP-hard in general, it is common to replace the matrix rank with the nuclear norm, which acts as a convenient convex surrogate. While elegant theoretical conditions elucidate when this replacement is likely to be successful, they are highly restrictive and convex algorithms fail when the ambient rank is too high or when the constraint set is poorly structured. Non-convex alternatives fare somewhat better when carefully tuned; however, convergence to locally optimal solutions remains a continuing source of failure. Against this backdrop we derive a deceptively simple and parameter-free probabilistic PCA-like algorithm that is capable, over a wide battery of empirical tests, of successful recovery even at the theoretical limit where the number of measurements equals the degrees of freedom in the unknown low-rank matrix. Somewhat surprisingly, this is possible even when the affine constraint set is highly ill-conditioned. While proving general recovery guarantees remains evasive for non-convex algorithms, Bayesian-inspired or otherwise, we nonetheless show conditions whereby the underlying cost function has a unique stationary point located at the global optimum; no existing cost function we are aware of satisfies this property. The algorithm has also been successfully deployed on a computer vision application involving image rectification and a standard collaborative filtering benchmark.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

1. Introduction

Recently there has been a surge of interest in finding minimum rank matrices subject to some problem-specific constraints often characterized as an affine set (Candès et al., 2011; Candès & Recht, 2009; Hu et al., 2013; Liu et al., 2013; Lu et al., 2014; Mohan & Fazel, 2012; Zhang et al., 2012). Mathematically this involves solving

$$\min_{\boldsymbol{X}} \ \text{rank}[\boldsymbol{X}] \quad \text{s.t. } \boldsymbol{b} = \mathcal{A}(\boldsymbol{X}), \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ is the unknown matrix, $\boldsymbol{b} \in \mathbb{R}^p$ represents a vector of observations and $\mathcal{A} : \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^p$ denotes a linear mapping. An important special case of (1) commonly applied to collaborative filtering is the matrix completion problem

$$\min_{\boldsymbol{X}} \ \text{rank}[\boldsymbol{X}] \quad \text{s.t. } \boldsymbol{X}_{ij} = (\boldsymbol{X}_0)_{ij}, \ (i,j) \in \Omega, \tag{2}$$

where $\boldsymbol{X}_0$ is a low-rank matrix we would like to recover, but we are only able to observe elements from the set Ω (Candès & Recht, 2009; Hu et al., 2013). Unfortunately however, both this special case and the general problem (1) are well-known to be NP-hard, and the rank penalty itself is non-smooth. Consequently, a popular alternative is to instead compute

$$\min_{\boldsymbol{X}} \ \sum_i f(\sigma_i[\boldsymbol{X}]) \quad \text{s.t. } \boldsymbol{b} = \mathcal{A}(\boldsymbol{X}), \tag{3}$$

where $\sigma_i[\boldsymbol{X}]$ denotes the i-th singular value of $\boldsymbol{X}$ and f is usually a concave, non-decreasing function (or nearly so). In the special case where $f(z) = I[z \neq 0]$ (i.e., an indicator function) we retrieve the matrix rank; however, smoother surrogates such as $f(z) = \log z$ or $f(z) = z^q$ with $q \leq 1$ are generally preferred for optimization purposes. When $f(z) = z$, (3) reduces to convex nuclear norm minimization. A variety of celebrated theoretical results have quantified specific conditions, heavily dependent on the singular values of matrices in the nullspace of $\mathcal{A}$, where the minimum nuclear norm solution is guaranteed to coincide with

that of minimal rank (Candès et al., 2011; Candès & Recht, 2009; Liu et al., 2013; Mohan & Fazel, 2012). However, these guarantees typically only apply to a highly restrictive set of rank minimization problems, and in a practical setting non-convex algorithms can succeed in a much broader range of conditions (Hu et al., 2013; Lu et al., 2014; Mohan & Fazel, 2012).

In Section 2 we will summarize state-of-the-art non-convex rank minimization algorithms that operate under affine constraints and point out some of their shortcomings. This will be followed in Section 3 by the derivation of an alternative approach using Bayesian modeling techniques adapted from probabilistic PCA (Tipping & Bishop, 1999). Section 4 will then describe properties of global and local solutions as well as special cases where any stationary point is guaranteed to have optimal rank, illustrating the intrinsic underlying smoothing mechanism which leads to success over competing methods. Finally, Section 5 contains a wide variety of numerical comparisons that highlight the efficacy of our algorithm. An extended journal version (Xin & Wipf, 2014) provides technical proofs, illustrations, and additional experiments as well as a computer vision application involving image rectification and a standard collaborative filtering benchmark. Before proceeding, we summarize two main contributions as follows:

- Bayesian inspiration can take uncountably many different forms and parameterizations, but the devil is in the details and existing methods offer little opportunity for both theoretical inquiry and substantial performance gains solving (1). In this regard, we apply carefully-tailored modifications to a veteran probabilistic PCA model leading to systematic theoretical and empirical insights and advantages.
- Over a wide battery of controlled experiments with ground-truth data, our approach outperforms all exising algorithms that we are aware of, Bayesian or otherwise; this includes direct head-to-head comparisons using the exact experimental designs and code prepared by original authors. In fact, even when $\mathcal{A}$ is ill-conditioned we are consistently able to solve (1) right up to the theoretical limit of any possible algorithm, which has never been demonstrated previously.

2. Related Work

Here we focus on a few of the latest and most effective rank minimization algorithms, all developed within the last few years and evaluated favorably against the state-of-the-art. In the non-convex regime, effective optimization strategies attempt to at least locally minimize (3), often exceeding the performance of the convex nuclear norm. For example, (Mohan & Fazel, 2012) derives a family of *iterative reweighted least squares* (IRLS) algorithms applied to $f(z) = (z^2 + \gamma)^{q/2}$ with $q, \gamma > 0$ as tuning parameters. A related penalty also considered, which coincides with the limit as $q \to 0$ (up to an inconsequential scaling and translation), is $f(z) = \log(z^2 + \gamma)$. This case also maintains an intimate connection with rank given that $\log z = \lim_{q \to 0} q^{-1}(z^q - 1) \equiv I[z \neq 0]$. Consequently, when γ is small, $\sum_i \log(\sigma_i[\boldsymbol{X}]^2 + \gamma)$ behaves much like the rank, albeit with nonzero gradients away from zero.

The IRLS0 algorithm from (Mohan & Fazel, 2012) represents the best-performing special case of the above, where $\sum_i \log(\sigma_i[\boldsymbol{X}]^2 + \gamma)$ is minimized using a homotopy continuation scheme merged with IRLS. Here a fixed γ is replaced with a decreasing sequence $\{\gamma^k\}$, the rationale being that when γ^k is large, the cost function is relatively smooth and devoid of local minima. As the iterations k progress, γ^k is reduced, and the cost behaves more like the matrix rank function. However, because now we are more likely to be within a reasonably good basin of attraction, spurious local minima are more easily avoided. The downside of this procedure is that it requires a pre-defined heuristic for reducing γ^k, and this schedule may be problem specific. Moreover, there is no guarantee that a global solution will ever be found.

In a related vein, (Lu et al., 2014) derives a family of *iterative reweighted nuclear norm* (IRNN) algorithms that can be applied to virtually any concave non-decreasing function f, even when f is non-smooth, unlike IRLS. For effective performance however the authors suggest a continuation strategy similar to IRLS0. Moreover, additional tuning parameters are required for different classes of functions f and it remains unclear which choices are optimal. While the reported results are substantially better than when using the convex nuclear norm, in our experiments IRLS0 seems to perform slightly better, possibly because the quadratic least squares inner loop is less aggressive in the initial stages of optimization than weighted nuclear norm minimization, leading to a better overall trajectory. Regardless, all of these affine rank minimization algorithms fail well before the theoretical recovery limit is reached, when the number of observations p equals the number of degrees of freedom in the low-rank matrix we wish to recover. Specifically, for an $n \times m$, rank r matrix, the number of degrees of freedom is given by $r(m+n)-r^2$, hence $p = r(m+n)-r^2$ is the best-case boundary. In practice if $\mathcal{A}$ is ill-conditioned or degenerate the achievable limit may be more modest.

A third approach relies on replacing the convex nuclear norm with a truncated non-convex surrogate (Hu et al., 2013). While some competitive results for image impainting via matrix completion are shown, in practice the proposed algorithm has many parameters to be tuned via cross-validation. Moreover, recent comparisons contained in (Lu et al., 2014) show that default settings perform rela-

tively poorly. Additionally, non-convex algorithms can be derived using a straightforward application of alternating minimization (Jain et al., 2013). The basic idea is to assume $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^T$ for some low-rank matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ and then solve $\min_{\boldsymbol{U},\boldsymbol{V}} \|b - \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^T)\|_{\mathcal{F}}$ via coordinate decent. The downside of this approach is that it can be sensitive to data correlations and requires that $\boldsymbol{U}$ and $\boldsymbol{V}$ be parameterized with the correct rank. In contrast, our emphasis here is on algorithms that require no prior knowledge whatsoever regarding the true rank. This is especially important in application extensions that may manage multiple low-rank matrices such that prior knowledge of all individual ranks is not feasible.

From a probabilistic perspective, previous work has applied Bayesian formalisms to rank minimization problems, although not specifically within an affine constraint set. For example, (Babacan et al., 2012; Ding et al., 2011; Wipf, 2012) derive robust PCA algorithms built upon the linear summation of a rank penalty and an element-wise sparsity penalty. While (Babacan et al., 2012) does consider the special case of matrix completion, none of these algorithms have been augmented and rigorously analyzed in the context of rank minimization with general affine constraints. Moreover, the limited analysis that does exist in (Wipf, 2012) actually just follows from the element-wise sparsity component intrinsic to robust PCA, without which the model effectively reduces to regular PCA devoid of any theoretical uncertainty. So the general affine constraints really are the key differentiating factor. Finally then, from a motivational standpoint, the basic probabilistic starting point we will adopt can be viewed as a careful re-parameterized generalization of the probabilistic PCA model from (Tipping & Bishop, 1999).

3. Alternative Algorithm Derivation

In contrast to the majority of existing algorithms organized around practical solutions to (3), here we adopt an alternative, probabilistic starting point. We first define the Gaussian likelihood function

$$p(\boldsymbol{b}|\boldsymbol{X};\mathcal{A},\lambda) \;\propto\; \exp\left[-\tfrac{1}{2\lambda}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{b}\|_2^2\right], \quad (4)$$

noting that in the limit as $\lambda \to 0$ this will enforce the same constraint set as in (1). Next we define an independent, zero-mean Gaussian prior distribution with covariance $\nu_i\boldsymbol{\Psi}$ on each column of $\boldsymbol{X}$, denoted $\boldsymbol{x}_i$ for all $i = 1,\ldots,m$. This produces the aggregate prior on $\boldsymbol{X}$ given by

$$p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) = \prod_i \mathcal{N}\left(\boldsymbol{x}_i;\boldsymbol{0},\nu_i\boldsymbol{\Psi}\right) \propto \exp\left[-\tfrac{1}{2}\boldsymbol{x}^\top\bar{\boldsymbol{\Psi}}^{-1}\boldsymbol{x}\right], \quad (5)$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is a positive semi-definite symmetric matrix,[1] $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_m]^\top$ is a non-negative vector, $\boldsymbol{x} = \text{vec}[\boldsymbol{X}]$ (column-wise vectorization), and $\bar{\boldsymbol{\Psi}} = \text{diag}[\boldsymbol{\nu}] \otimes \boldsymbol{\Psi}$, with $\otimes$ denoting the Kronecker product. It is important to stress here that we do not necessarily believe that the unknown $\boldsymbol{X}$ actually follows such a Gaussian distribution per se. Rather, we adopt (5) primarily because it will lead to an objective function with desirable properties related to solving (1). Moving forward, given both likelihood and prior are Gaussian, the posterior $p(\boldsymbol{X}|\boldsymbol{b};\boldsymbol{\Psi},\boldsymbol{\nu},\mathcal{A},\lambda)$ is also Gaussian, with mean given by an $\hat{\boldsymbol{X}}$ such that

$$\hat{\boldsymbol{x}} = \text{vec}[\hat{\boldsymbol{X}}] = \bar{\boldsymbol{\Psi}}\boldsymbol{A}^\top\left(\lambda\boldsymbol{I} + \boldsymbol{A}\bar{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{b}. \quad (6)$$

Here $\boldsymbol{A} \in \mathbb{R}^{p\times nm}$ is a matrix defining the linear operator $\mathcal{A}$ such that $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}$ reproduces the feasible region in (1). From this expression it is clear that, if $\boldsymbol{\Psi}$ represents a low-rank covariance matrix, then each column of $\hat{\boldsymbol{X}}$ will be constrained to a low-dimensional subspace resulting overall in a low-rank estimate as desired. Of course for this simple strategy to be successful we require some way of determining a viable $\boldsymbol{\Psi}$ and the scaling vector $\boldsymbol{\nu}$.

A common Bayesian strategy in this regard is to marginalize over $\boldsymbol{X}$ and then maximize the resulting likelihood function with respect to $\boldsymbol{\Psi}$ and $\boldsymbol{\nu}$ (Tipping, 2001; Wipf, 2012; Wipf et al., 2011). This involves solving

$$\max_{\boldsymbol{\Psi}\in H^+,\boldsymbol{\nu}\geq 0} \int p(\boldsymbol{b}|\boldsymbol{X};\mathcal{A},\lambda)p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu})d\boldsymbol{X}, \quad (7)$$

where H^+ denotes the set of positive semi-definite and symmetric $n \times n$ matrices. After a $-2\log$ transformation, this is equivalent to minimizing the cost function

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\Psi},\boldsymbol{\nu}) = \boldsymbol{b}^\top\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b} + \log|\boldsymbol{\Sigma}_b|, \\ &\boldsymbol{\Sigma}_b = \boldsymbol{A}\bar{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda\boldsymbol{I}, \quad \bar{\boldsymbol{\Psi}} = \text{diag}[\boldsymbol{\nu}] \otimes \boldsymbol{\Psi}, \end{aligned} \quad (8)$$

where $\boldsymbol{\Sigma}_b$ is the covariance of $\boldsymbol{b}$ given $\boldsymbol{\Psi}$ and $\boldsymbol{\nu}$. Minimizing (8) is a non-convex optimization problem, and we employ standard upper bounds for this purpose leading to an EM-like algorithm somewhat related to (Tipping & Bishop, 1999). In particular, we compute separate bounds, parameterized by auxiliary variables, for both the first and second terms of $\mathcal{L}(\boldsymbol{\Psi},\boldsymbol{\nu})$. While the general case can easily be handled and may be applicable for more challenging problems, here for simplicity and ease of presentation we consider minimizing $\mathcal{L}(\boldsymbol{\Psi}) \triangleq \mathcal{L}(\boldsymbol{\Psi},\boldsymbol{\nu} = \boldsymbol{1})$, meaning all

[1] Technically $\boldsymbol{\Psi}$ must be positive definite for the inverse in (5) to be defined. However, we can accommodate the semi-definite case using the following convention. Without loss of generality assume that $\bar{\boldsymbol{\Psi}} = \boldsymbol{R}\boldsymbol{R}^\top$ for some matrix $\boldsymbol{R}$. We then qualify that $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) = 0$ if $\boldsymbol{x} \notin \text{span}[\boldsymbol{R}]$, and $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}\right]$ otherwise. Equivalently, for convenience (and with slight abuse of notation) we define $\boldsymbol{x}^\top\bar{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \infty$ when $\boldsymbol{x} \notin \text{span}[\boldsymbol{R}]$, and $\boldsymbol{x}^\top\bar{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}$ otherwise. This will come in handy, for example, when interpreting the bound in (9) below. Note also that the final cost function (8) we will ultimately be minimizing requires no such inverse anyway.

elements of $\boldsymbol{\nu}$ are fixed at one (and such is the case for all experiments reported herein, although we are currently exploring situations where this added generality could be especially helpful).

Based on (Wipf et al., 2011), for the first term in (8), we have

$$\boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \;\; \leq \;\; \tfrac{1}{\lambda}\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \boldsymbol{x}^\top \bar{\boldsymbol{\Psi}}^{-1} \boldsymbol{x} \tag{9}$$

with equality whenever $\boldsymbol{x}$ satisfies (6). For the second term we use

$$\begin{aligned} \log|\boldsymbol{\Sigma}_b| &\equiv m \log|\boldsymbol{\Psi}| + \log\left|\lambda \boldsymbol{A}^\top \boldsymbol{A} + \bar{\boldsymbol{\Psi}}^{-1}\right| \\ &\leq m \log|\boldsymbol{\Psi}| + \mathrm{tr}\left[\boldsymbol{\Psi}^{-1} \nabla_{\Psi^{-1}}\right] + C, \end{aligned} \tag{10}$$

where because $\log\left|\lambda \boldsymbol{A}^\top \boldsymbol{A} + \bar{\boldsymbol{\Psi}}^{-1}\right|$ is concave with respect to $\boldsymbol{\Psi}^{-1}$, we can upper bound it using a first-order approximation with a bias term C that is independent of $\boldsymbol{\Psi}$. Equality is obtained when the gradient satisfies

$$\nabla_{\Psi^{-1}} = \sum_{i=1}^{m} \boldsymbol{\Psi} - \boldsymbol{\Psi}\boldsymbol{A}_i^\top \left(\boldsymbol{A}\bar{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_i \boldsymbol{\Psi}, \tag{11}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{p \times n}$ is defined such that $\boldsymbol{A} = [\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m]$. Finally given the upper bounds from (9) and (10) with $\boldsymbol{X}$ and $\nabla_{\Psi^{-1}}$ fixed, we can compute the optimal $\boldsymbol{\Psi}$ in closed form by optimizing the relevant $\boldsymbol{\Psi}$-dependent terms via

$$\begin{aligned} \boldsymbol{\Psi}^{opt} &= \arg\min_{\boldsymbol{X}} \mathrm{tr}\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{X}\boldsymbol{X}^\top + \nabla_{\Psi^{-1}}\right)\right] + m\log|\boldsymbol{\Psi}| \\ &= \tfrac{1}{m}\left[\hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^\top + \nabla_{\Psi^{-1}}\right]. \end{aligned} \tag{12}$$

By iteratively computing (6), (11), and (12), we can then obtain an estimate for $\boldsymbol{\Psi}$, and more importantly, a corresponding estimate for $\boldsymbol{X}$ given by (6) at convergence. We refer to this basic procedure as BARM for *Bayesian Affine Rank Minimization*. The next section will describe in detail why it is particularly well-suited for solving problems such as (1), as well as additional algorithmic enhancements.

4. Properties of BARM

As discussed in Section 2 one nice property of the $\sum_i \log\left(\sigma_i[\boldsymbol{X}]\right)$ penalty employed (approximately) by IRLS0 (Mohan & Fazel, 2012) is that it can be viewed as a smooth version of the matrix rank function while still possessing the same set of minimum, both global and local, over the affine constraint set, at least if we consider the limiting situation of $\sum_i \log\left(\sigma_i[\boldsymbol{X}]^2 + \gamma\right)$ when γ becomes small so that we may avoid the distracting singularity of $\log 0$. Additionally, it possesses an attractive form of scale invariance, meaning that if $\boldsymbol{X}^*$ is an optimal feasible solution, a block-diagonal rescaling of $\boldsymbol{A}$ nevertheless leads to an equivalent rescaling of the optimum (without the need for solving an additional optimization problem using the new $\boldsymbol{A}$). This is very much unlike the nuclear norm or other non-convex surrogates that penalize the singular values of $\boldsymbol{X}$ in a scale-dependent manner.

In contrast, the proposed algorithm is based on a very different Gaussian statistical model with seemingly a more tenuous connection with rank minimization. Encouragingly however, the proposed cost function enjoys the same global/local minima properties as $\sum_i \log\left(\sigma_i[\boldsymbol{X}]^2 + \gamma\right)$ with $\gamma \to 0$. Before presenting these results, we define $\mathrm{spark}[\boldsymbol{A}]$ as the smallest number of linearly dependent columns in matrix $\boldsymbol{A}$ (Donoho & Elad, 2003).

Lemma 1. *Let $\boldsymbol{b} = \boldsymbol{A}\,\mathrm{vec}[\boldsymbol{X}]$, where $\boldsymbol{A} \in \mathbb{R}^{p \times nm}$ satisfies $\mathrm{spark}[\boldsymbol{A}] = p + 1$. Also define r as the smallest rank of any feasible solution. Then if $r < p/m$, any global minimizer $\{\boldsymbol{\Psi}^*, \boldsymbol{\nu}^*\}$ of (8) in the limit $\lambda \to 0$ is such that $\boldsymbol{x}^* = \bar{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top\left(\boldsymbol{A}\bar{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top\right)^\dagger \boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*] = r$ with $\mathrm{vec}[\boldsymbol{X}^*] = \boldsymbol{x}^*$.*

Lemma 2. *Additionally, let $\tilde{\boldsymbol{A}} = \boldsymbol{A}\boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{diag}[\alpha_1\boldsymbol{\Gamma}, \ldots, \alpha_m\boldsymbol{\Gamma}]$ is a block-diagonal matrix with invertible blocks $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times n}$ of unit norm scaled with coefficients $\alpha_i > 0$. Then iff $\{\boldsymbol{\Psi}^*, \boldsymbol{\nu}^*\}$ is a minimizer (global or local) to (8) in the limit $\lambda \to 0$, then $\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}^*, \mathrm{diag}[\boldsymbol{\alpha}]^{-1}\boldsymbol{\nu}^*\}$ is a minimizer when $\tilde{\boldsymbol{A}}$ replaces $\boldsymbol{A}$. The corresponding estimates of $\boldsymbol{X}$ are likewise in one-to-one correspondence.*

Remarks: The assumption $r = \mathrm{rank}[\boldsymbol{X}^*] < p/m$ in Lemma 1 is completely unrestrictive, especially given that a unique, minimal-rank solution is only theoretically possible by *any* algorithm if $p \geq (n+m)r - r^2$, which is much more restrictive than $p > rm$. Hence the bound we require is well above that required for uniqueness anyway. Likewise the spark assumption will be satisfied for any $\boldsymbol{A}$ with even an infinitesimal (continuous) random component. Consequently, we are essentially always guaranteed that BARM possesses the same global optimum as the rank function. Regarding Lemma 2, no surrogate rank penalty of the form $\sum_i f(\sigma_i[\boldsymbol{X}])$ can achieve this result except for $f(z) = \log z$, or inconsequential limiting translations and rescalings of the log such as the indicator function $I[z \neq 0]$ (which is related to the log via arguments in Section 2).

While these results are certainly a useful starting point, the real advantage of adopting the BARM cost function is that locally minimizing solutions are exceedingly rare, largely as a consequence of the marginalization process in (7), and in some cases provably so. A specialized example of this smoothing can be quantified in the following scenario.

Suppose $\boldsymbol{A}$ is now block diagonal, with diagonal blocks $\boldsymbol{A}_i$ such that $\boldsymbol{b}_i = \boldsymbol{A}_i\boldsymbol{x}_i$ producing the aggregate observation vector $\boldsymbol{b} = [\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_m^\top]^\top$. While somewhat restricted,

this situation nonetheless includes many important special cases, including canonical matrix completion and generalized matrix completion where elements of $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{X}_0$ are observed instead of $\boldsymbol{X}_0$ directly.

Theorem 1. *Let $\boldsymbol{b} = \boldsymbol{A}\,\mathrm{vec}[\boldsymbol{X}]$, where $\boldsymbol{A}$ is block diagonal, with blocks $\boldsymbol{A}_i \in \mathbb{R}^{p_i \times n}$. Moreover, assume $p_i > 1$ for all i and that $\cap_i null[\boldsymbol{A}_i] = \emptyset$. Then if $\min_{\boldsymbol{X}} \mathrm{rank}[\boldsymbol{X}] = 1$ in the feasible region, any minimizer $\{\boldsymbol{\Psi}^*, \boldsymbol{\nu}^*\}$ of (8) (global or local) in the limit $\lambda \rightarrow 0$ is such that $\boldsymbol{x}^* = \bar{\boldsymbol{\Psi}}^* \boldsymbol{A}^\top \left(\boldsymbol{A}\bar{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top\right)^\dagger \boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*] = 1$ with $\mathrm{vec}[\boldsymbol{X}^*] = \boldsymbol{x}^*$. Furthermore, no cost function in the form of (3) can satisfy the same result. In particular, there can always exist local and/or global minima with rank greater than one.*

Remarks: This result implies that, under extremely mild conditions, which do not even depend on the concentration properties of $\boldsymbol{A}$, the proposed cost function has no minima that are not global minima. (The minor technical condition regarding nullspace intersections merely ensures that high-rank components cannot simultaneously "hide" in the nullspace of every measurement matrix $\boldsymbol{A}_i$; the actual $\boldsymbol{A}$ operator may still be highly ill-conditioned.) Thus any algorithm with provable convergence to some local minimizer is guaranteed to obtain a globally optimal solution.[2] Moreover, such a guarantee is not possible with any other penalty function of the standard form $\sum_i f(\sigma_i[\boldsymbol{X}])$, which is the typical recipe for rank minimization algorithms, convex or not. Additionally, if a unique rank-one solution exists to (1), then the unique minimizing solution to (8) will produce this $\boldsymbol{X}$ via (6). Crucially, this will occur even when the minimal number of measurements $p = n + m - 1$ are available, unlike any other algorithm we are aware of that is blind to the true underlying rank.[3] And the underlying intuition, that local minima are smoothed away, nonetheless carries over to situations where the rank is greater than one. An enlightening visualization of this smoothing effect can be found in (Xin & Wipf, 2014).

Convergence: Previous results are limited to exploring aspects of the underlying BARM cost function. Regarding the BARM algorithm itself, by construction the updates generated by (6), (11), and (12) are guaranteed to reduce or leave unchanged $\mathcal{L}(\boldsymbol{\Psi})$ at each iteration. However, this is not technically sufficient to guarantee convergence to a stationary point of the cost function unless the additional conditions of Zangwill's Global Convergence Theorem are satisfied (Zangwill, 1969). However, provided we add a small regularization factor $\gamma\,\mathrm{tr}[\boldsymbol{\Psi}^{-1}]$, with $\gamma > 0$, then it can be shown that any cluster point of the resulting sequence of iterations $\{\boldsymbol{\Psi}^k\}$ must be a stationary point. Moreover, because the sequence is bounded, there will always exist at least one cluster point, and therefore the algorithm is guaranteed to at least converge to a set of parameter values $\mathcal{S}$ such that for any $\boldsymbol{\Psi}^* \in \mathcal{S}$, $\mathcal{L}(\boldsymbol{\Psi}^*) + \gamma\,\mathrm{tr}[(\boldsymbol{\Psi}^*)^{-1}]$ is a stationary point. Finally, we should mention that this extra γ factor is akin to the homotopy continuation regularizer used by the IRLS0 algorithm (Mohan & Fazel, 2012) as discussed in Section 2. However, whereas IRLS0 requires a carefully-chosen, decreasing sequence $\{\gamma^k\}$ with $\gamma^k > 0$ *both* to prove convergence *and* to avoid local minimum (and without this factor the algorithm performs very poorly in practice), for BARM a small, fixed factor only need be included as a technical necessity for proving formal convergence; in practice it can be fixed to exactly zero.

Symmetrization Improvements: Despite the promising theoretical attributes of BARM, there remains one important artifact of its probabilistic origins not found in more conventional existing rank minimization algorithms. In particular, other algorithms rely upon a symmetric penalty function that is independent of whether we are working with $\boldsymbol{X}$ or $\boldsymbol{X}^\top$. All methods that reduce to (3) fall into this category, e.g., nuclear norm minimization, IRNN, or IRLS0. In contrast, our method relies on defining a distribution with respect to the columns of $\boldsymbol{X}$. Consequently the underlying cost function is not identical when derived with respect to $\boldsymbol{X}$ or $\boldsymbol{X}^\top$, a difference which will depend on $\boldsymbol{A}$. While globally optimal solutions should nonetheless be the same, the convergence trajectory could depend on this distinction leading to different local minima in certain circumstances. Although either construction leads to low-rank solutions, we may nonetheless expect improvement if we can somehow symmetrize the algorithm formulation. To accomplish this, we consider a Gaussian prior on $\boldsymbol{x} = \mathrm{vec}[\boldsymbol{X}]$ with a covariance formed using a block-wise averaging of covariances defined over rows and columns, denoted $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_c$ respectively. The overall covariance is then given by the Kronecker sum $\bar{\boldsymbol{\Psi}} = 1/2\,(\boldsymbol{\Psi}_r \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{\Psi}_c)$. The estimation process then proceeds in a similar fashion as before but with modifications and alternate upper-bounds that accommodate for this merger. For reported experimental results this symmetric version of BARM is used, with complete update rules and a discussion of computational complexity deferred to (Xin & Wipf, 2014).

5. Experimental Validation

This section compares BARM with existing state-of-the-art affine rank minimization algorithms. For BARM, in

[2] Note also that with minimal additional effort, it can be shown that no suboptimal stationary points of any kind, including saddle points, are possible.

[3] It is important to emphasize that the difficulty of estimating the optimal low-rank solution is based on the ratio of the d.o.f. in $\boldsymbol{X}$ to the number of observations p. Consequently, estimating $\boldsymbol{X}$ even with r small can be challenging when p is also small, meaning $\boldsymbol{A}$ is highly overcomplete.

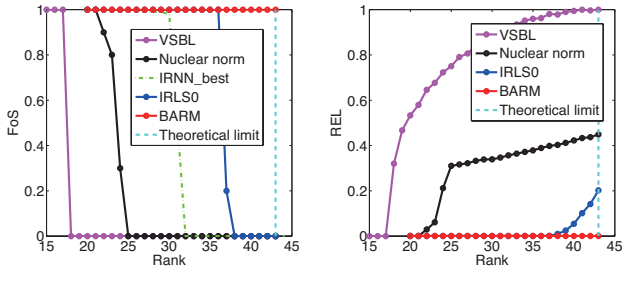


Figure 1. Matrix completion comparisons (avg of 10 trials)

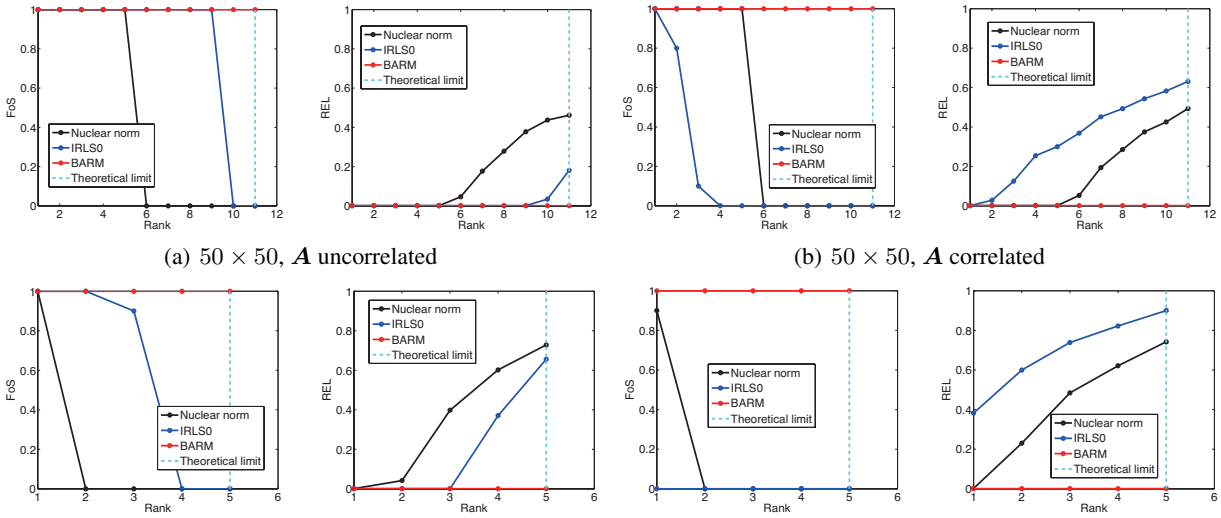
all cases we simply used $\lambda = 10^{-10}$ (effectively zero), and hence no tuning parameters are required. Likewise, nuclear norm minimization (Candès & Recht, 2009; Zhang et al., 2012) requires no tuning parameters beyond implementation-dependent control parameters frequently used to enhance convergence speed (however the global minimum is unaltered given that the problem is convex). For the IRLS0 algorithm, we used our own implementation as the algorithm is straightforward and no code was available for the case of general $\mathcal{A}$; we based the required decreasing γ_k sequence on suggestions from (Mohan & Fazel, 2012). IRLS0 code is available from the original authors for matrix completion; however, the results obtained with this code are not better than those obtained with our version. For the IRNN algorithm, we did not have access to code for general $\mathcal{A}$, nor specific details of how various parameters should be set in the general case. Note also that IRNN has multiple parameters to tune even in noiseless problems unlike BARM. Therefore we report results directly from (Lu et al., 2014) where available. Note that both (Lu et al., 2014) and (Mohan & Fazel, 2012) show superior results to a number of other algorithms; we do not generally compare with these others given that they are likely no longer state-of-the-art and may clutter the presentation.

As stated previously, our focus here is on algorithms that do not require knowledge of the true rank of the optimal solution, and hence we do not include comparisons with (Jain et al., 2013) or the normalized hard thresholding algorithm from (Tanner & Wei, 2013). Regardless, we have nonetheless conducted numerous experiments with these algorithms, and even when the correct rank is provided, results are inferior to BARM, especially when correlated measurements are used (see (Xin & Wipf, 2014)). However, we do show limited empirical results with the variational sparse Bayesian algorithm (VSBL) from (Babacan et al., 2012) because of its Bayesian origins, although the underlying parameterization is decidedly different from BARM. But these results are limited to matrix completion as VSBL presently does not handle general affine constraints. Results from VSBL were obtained using publicly available code from the authors.

Matrix Completion: We begin with the matrix completion problem from (2), in part because this allows us to compare our results with the latest algorithms even when code is not available. For this purpose we reproduce the exact same experiment from (Lu et al., 2014), where a rank r matrix is generated as $\boldsymbol{X}_0 = \mathbf{M}_L\mathbf{M}_R$, with $\mathbf{M}_L \in \mathbb{R}^{n\times r}$ and $\mathbf{M}_R \in \mathbb{R}^{r\times m}$ ($n = m = 150$) as iid $\mathcal{N}(0,1)$ random matrices. 50% of all entries are then hidden uniformly at random. The *relative error* (REL) given by $\|\boldsymbol{X}_0 - \hat{\boldsymbol{X}}\|_{\mathcal{F}}/\|\boldsymbol{X}_0\|_{\mathcal{F}}$ is computed for each trial and averaged as r is varied. Likewise, we compute the *frequency of success* (FoS) score which measures the percentage of trials where the REL is below 10^{-3}. Results are shown in Figure 1 where BARM is the only algorithm capable of reaching the theoretical recovery limit, beyond which $p = 0.5 \times 150^2 = 11250$ is surpassed by the number of degrees of freedom in $\boldsymbol{X}_0$, in this case $2 \times 150 \times 44 - 44^2 = 11264$. Note that FoS values were reported in (Lu et al., 2014) over a wide range of non-convex IRNN algorithms. The green curve represents the best performing candidate from this pool as tuned by the original authors; REL values were unavailable. Interestingly, although VSBL is based on a somewhat related probabilistic model to BARM, the underlying parameterization, cost function, and update rules are entirely different and do not benefit from strong theoretical underpinnings. Hence performance does not always match recent state-of-the-art algorithms, although from a computational standpoint it is quite efficient.

Besides BARM, the IRLS0 algorithm also displayed better performance than the other methods. This motivated us to reproduce some of the matrix completion experiments from (Mohan & Fazel, 2012) so as to provide direct head-to-head comparisons with the authors' original implementation. For this purpose, $\boldsymbol{X}_0$ is conveniently generated in the same way as above; however, values of n, m, r, and the percentage of missing entries are varied while evaluating reconstructions using FoS. While (Mohan & Fazel, 2012) tests a variety of combinations of these values to explore varying degrees of problem difficulty, here we only reproduce the most challenging cases to see if BARM is still able to produce superior reconstruction accuracy. In this respect problem difficulty is measured by the *degrees of freedom ratio* (FR) given by FR$= r(n+m-r)/p$ as defined in (Mohan & Fazel, 2012). We also only include experiments where algorithms are blind to the true rank of $\boldsymbol{X}_0$.[4] Results are shown in Table 1, where we have also displayed the published results of three additional algorithms that were compared with IRLS0 in (Mohan & Fazel, 2012), namely, IHT (Jain et al., 2010), FPCA (Goldfarb & Ma, 2011) and Optspace (Keshavan & Oh, 2009). From the table we observe that, in the most difficult problem considered in (Mohan & Fazel, 2012), IRLS0 achieved only a 0.5 FoS score (meaning failure 50% of the time) while BARM

[4]Note that IRLS0 can be modified to account for the true rank if such knowledge were available.



(a) 50×50, $\boldsymbol{A}$ uncorrelated (b) 50×50, $\boldsymbol{A}$ correlated

(c) 100×100, $\boldsymbol{A}$ uncorrelated (d) 100×100, $\boldsymbol{A}$ correlated

Figure 2. Comparisons with general affine constraints (avg of 10 trials)

still achieves a perfect 1.0.

General A: Next we consider the more challenging problem involving arbitrary affine constraints. The desired low-rank $\mathbf{X}_0$ is generated in the same way as above. We then consider two types of linear mappings where $\boldsymbol{A}$ is generated as: (i) an iid $\mathcal{N}(0,1)$, $p \times n^2$ matrix, and (ii) $\sum_{i=1}^{p} i^{-1/2} \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $\boldsymbol{u}_i \in \mathbb{R}^p$ and $\boldsymbol{v}_i \in \mathbb{R}^{n^2}$ are iid $\mathcal{N}(0,1)$ vectors. The latter is meant to explore less-than-ideal conditions where the linear operator displays correlations and may be somewhat ill-conditioned. Figure 2 displays aggregate results when $\boldsymbol{X}_0$ is 50×50 and 100×100. In both cases $p = 1000$ observations are used, and therefore the corresponding measurement matrices $\boldsymbol{A}$ are 1000×2500 and 1000×10000 respectively. We then vary r from 1 up to the theoretical limit corresponding to problem size. Again we observe that BARM is consistently able to work up to the limit, even when the $\boldsymbol{A}$ operator is no longer an ideal Gaussian. In general, we have explored a wide range of empirical conditions too lengthly to report here, and it is only very rarely, and always near the theoretical boundary, where BARM occasionally may not succeed. We explore such failure cases in the next section.

Failure Case Analysis: Thus far we have not shown any cases where BARM actually fails. Of course solving (1) for general a $\boldsymbol{A}$ is NP-hard so recovery failures certainly must exist in some circumstances when using a polynomial-time algorithm such as BARM. Although we certainly cannot explore every possible scenario, it behooves us to probe more carefully for conditions under which such errors may occur. One way to accomplish this is to push the problem difficulty even further towards the theoretical limit by reducing the number of measurements p as follows.

With the number of observations fixed at $p = 1000$ and a general measurement matrix $\boldsymbol{A}$, the previous section examined the recovery of 50×50 and 100×100 matrices as the rank was varied from 1 to the recovery limit ($r = 11$ for the 50×50 case; $r = 5$ for the 100×100 case). However, it is still possible to make the problem even more challenging by fixing r at the limit and then reducing p until it exactly equals the degrees of freedom $2n^2 - r^2$. With $\{n = 50, r = 11\}$ this occurs at $p = 979$, for $\{n = 100, r = 5\}$ this occurs at $p = 975$.

Table 1. Matrix completion comparisons of BARM with IRLS0 on the three hardest problems from (Mohan & Fazel, 2012). Published results in (Mohan & Fazel, 2012) included for comparison.

Problem			IRLS0	IHT	FPCA	Opts	BARM
FR	n(=m)	r	FoS	FoS	FoS	FoS	FoS
0.78	500	20	0.9	0	0	0	1
0.8	40	9	1	0	0.5	0	1
0.87	100	14	0.5	0	0	0	1

We examined the BARM algorithm under these conditions with 10 additional trials using the uncorrelated $\boldsymbol{A}$ for each problem size. Encouragingly, BARM was still 30% successful with $\{n = 50, r = 11\}$, and 40% successful with $\{n = 100, r = 5\}$. However, it is interesting to further examine the nature of these failure cases. We notice that, although the recovery was technically classified as a failure since the relative error (REL) was above the stated threshold, the estimated matrices are of almost exactly the correct minimal rank. Illustrations of the actual singular values can be found in (Xin & Wipf, 2014). Hence BARM has essentially uncovered an alternative solution with minimal rank that is nonetheless feasible by construction. We therefore speculate that right at the theoretical limit, when $\boldsymbol{A}$ is maximally overcomplete ($p \times n^2 = 979 \times 2500$ or 975×10000

Table 2. Further matrix completion comparisons of BARM with IRLS0 by reducing the number of measurements in the hardest problem from (Mohan & Fazel, 2012). Results with both FoS and FoRS metrics are reported (avg of 10 trials).

Problem			IRLSO		BARM	
FR	n(=m)	r	FoS	FoRS	FoS	FoRS
0.9	100	14	0	0	1	1
0.95	100	14	0	0	0.8	1
0.99	100	14	0	0	0.7	1

for the two problem sizes), there exists multiple feasible matrices with singular value spectral cut-off points indistinguishable from the optimal solution. Importantly, when the other algorithms we tested failed, the failure is much more dramatic and a clear spectral cut-off at the correct rank is not apparent.

This motivates a looser success criteria than FoS to account for the possibility of multiple (nearly) optimal solutions that may not necessarily be close with respect to relative error. For this purpose we define the *frequency of rank success* (FoRS) as the percentage of trials whereby a feasible solution $\hat{\boldsymbol{X}}$ is found such that $\sigma_r[\hat{\boldsymbol{X}}]/\sigma_{r+1}[\hat{\boldsymbol{X}}] > 10^3$, where $\sigma_i[\cdot]$ denotes the i-th singular value of a matrix and r is the rank of the true low-rank $\boldsymbol{X}_0$. In words, FoRS measures the percentage of trials such that roughly a rank r solution is recovered, regardless of proximity to $\boldsymbol{X}_0$.

Under this new criteria, all of the failure cases with respect to FoS described above, for both problem sizes, become successes; however, none of the other algorithms show improvement under this criteria, indicating that their original failures involved actual sub-optimal rank solutions. Something similar happens when we revisit the matrix completion experiments. For example, based on Table 1 the most difficult case involves FR$=0.87$; however, by further reducing p, we can push FR towards 1.0 to further investigate the break-down point of BARM. Results are shown in Table 2. While IRLS0 (which is the top performing algorithm in (Mohan & Fazel, 2012) and in our experiments besides BARM) fails 100% of the time via both metrics, BARM can achieve an FoS of 0.7 even when FR$=0.99$ and an FoRS of 1.0 in all cases.

We therefore adopt a more challenging measurement structure for $\boldsymbol{A}$ to better evaluate the limits of BARM performance to reveal potential failures by both FoS *and* FoRS metrics. Specifically, we first applied 2-D *discrete cosine transform* (DCT) to $\boldsymbol{X}_0$ and then randomly sampled p of the resulting DCT coefficients. Because both the DCT and the sampling sub-process are linear operations on the entries of $\boldsymbol{X}_0$, the whole process is representable via a matrix $\boldsymbol{A}$, which encodes highly structured information. Detailed results can be found in (Xin & Wipf, 2014). Two things stand out from this analysis. First, while the other algorithms display almost identical behavior under either metric, BARM failures under the FoS criteria are mostly converted to successes by the FoRS metric by recovering a matrix of near-optimal rank. Secondly, even though certain unequivocal failures emerge near the limits with this challenging DCT-based sampling matrix, BARM outperforms the other algorithms using either metric by a large margin.

To summarize, we have demonstrated that BARM is capable of recovering a low-rank matrix right up to the theoretical limit in a variety of scenarios using different types of measurement processes. Moreover, even in cases where it fails, it often nonetheless still produces a feasible $\hat{\boldsymbol{X}}$ with rank nearly identical to the generative low-rank $\boldsymbol{X}_0$, suggesting that multiple optimal solutions may be possible in challenging borderline cases. But when true unequivocal failures do occur, such failures tend to be near the theoretical boundary, and with greater likelihood when the dictionary displays significant structure (or correlations). While certainly we envision that, out of the infinite multitude of testing situations further significant pockets of BARM failure can be revealed, we nonetheless feel that BARM is quite promising relative to existing algorithms.

Noisy Simulations and Application Examples: Although our primary purpose has been to derive and rigorously analyze BARM, and Monte-Carlo experiments with noiseless ground-truth data are a convenient way to do this, we have also conducted both noisy simulation experiments, where BARM displays desirable stability, and application-specific tests; for space considerations these results are all deferred to (Xin & Wipf, 2014). For the applications, we consider two examples: image rectification and collaborative filtering for recommender systems. The former implicitly involves a general sampling operator $\boldsymbol{A}$, while the latter reduces to a standard matrix completion problem.

6. Conclusion

This paper explores a conceptually-simple, parameter-free algorithm for matrix rank minimization under affine constraints that is capable of successful recovery empirically observed to approach the theoretical limit over a broad class of experimental settings (including many not shown here) unlike any existing algorithms, and long after any convex guarantees break down. While our model is ultimately based on a Gaussian marginal likelihood function, variations of which have been analyzed thoroughly in the context of sparse estimation (Wipf et al., 2011), the affine rank minimization problem addressed here is considerably different. Moreover, our main theoretical result, Theorem 1, relies on a completely different underlying analysis technique; likewise for the symmeterization adaptation. This mirrors well-established differences between convex ℓ_1 and nuclear norm algorithms for compressive sensing and rank minimization respectively.

Acknowledgments

This work was supported in part by 973-2015CB351800, NSFC-61421062, 61210005.

References

Babacan, S Derin, Luessi, Martin, Molina, Rafael, and Katsaggelos, Aggelos K. Sparse bayesian methods for low-rank matrix estimation. *Signal Processing, IEEE Transactions on*, 60(8):3964–3977, 2012.

Candès, Emmanuel J and Recht, Benjamin. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Candès, Emmanuel J, Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Ding, Xinghao, He, Lihan, and Carin, Lawrence. Bayesian robust principal component analysis. *Image Processing, IEEE Transactions on*, 20(12):3419–3430, 2011.

Donoho, David L and Elad, Michael. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Goldfarb, Donald and Ma, Shiqian. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2): 183–210, 2011.

Hu, Yao, Zhang, Debing, Ye, Jieping, Li, Xuelong, and He, Xiaofei. Fast and accurate matrix completion via truncated nuclear norm regularization. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on*, 35 (9):2117–2130, 2013.

Jain, Prateek, Meka, Raghu, and Dhillon, Inderjit S. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pp. 937–945, 2010.

Jain, Prateek, Netrapalli, Praneeth, and Sanghavi, Sujay. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Keshavan, Raghunandan H and Oh, Sewoong. A gradient descent algorithm on the grassman manifold for matrix completion. *arXiv preprint arXiv:0910.5260*, 2009.

Liu, Guangcan, Lin, Zhouchen, Yan, Shuicheng, Sun, Ju, Yu, Yong, and Ma, Yi. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on*, 35(1):171–184, 2013.

Lu, Canyi, Tang, Jinhui, Yan, Shuicheng, and Lin, Zhouchen. Generalized nonconvex nonsmooth low-rank minimization. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*. IEEE, 2014.

Mohan, Karthik and Fazel, Maryam. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research (JMLR)*, 13(1):3441–3473, 2012.

Tanner, Jared and Wei, Ke. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

Tipping, Michael and Bishop, Christopher. Probabilistic principal component analysis. *J. Royal Statistical Society, Series B*, 61(3):611–622, 1999.

Tipping, Michael E. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research (JMLR)*, 1:211–244, 2001.

Wipf, David. Non-convex rank minimization via an empirical bayesian approach. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

Wipf, David P, Rao, Bhaskar D, and Nagarajan, Srikantan. Latent variable bayesian models for promoting sparsity. *Information Theory, IEEE Transactions on*, 57(9):6236–6255, 2011.

Xin, Bo and Wipf, David. Exploring algorithmic limits of matrix rank minimization under affine constraints. *arXiv preprint arXiv:1406.2504*, 2014.

Zangwill, Willard I. *Nonlinear programming: a unified approach*. Prentice Hall, 1969.

Zhang, Zhengdong, Ganesh, Arvind, Liang, Xiao, and Ma, Yi. Tilt: transform invariant low-rank textures. *International Journal of Computer Vision (IJCV)*, 99(1):1–24, 2012.