Exploring Algorithmic Limits of Matrix Rank Minimization Under Affine Constraints

Bo Xin, *Member, IEEE*, David Wipf, *Member, IEEE*, Yizhou Wang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***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. Nonconvex 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 nonconvex 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.**

***Index Terms*—Rank minimization, affine constraints, matrix completion, matrix recovery, empirical Bayes.**

I. 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 [1]–[7]. Mathematically this involves solving

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

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

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

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 [1], [2]. 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

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

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 [1], [3], [6]. However, these guarantees typically only apply to a highly restrictive set of rank minimization problems, and in a practical setting nonconvex algorithms can succeed in a much broader range of conditions [2], [5], [6].

In Section II 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 III by the derivation of an alternative approach using Bayesian modeling techniques adapted from probabilistic PCA [8]. Section IV will then describe connections with nuclear norm minimization, convergence issues, and properties of global and local solutions. The latter includes special cases whereby any stationary point of the intrinsic cost function is guaranteed to have optimal rank, illustrating an underlying smoothing mechanism which leads to success over competing methods. We next discuss algorithmic enhancements in Section V that further improve recovery performance in practice. Section VI contains a wide variety of numerical comparisons that highlight the efficacy of this algorithm, while Section Section VII presents a computer vision application involving image rectification and a standard collaborative filtering benchmark. Technical proofs and algorithm update rule details are contained in the Appendix. Portions of this work have previously appeared in conference proceedings [9].

Manuscript received October 22, 2014; revised March 9, 2015 and November 23, 2015; accepted February 26, 2016. Date of publication April 7, 2016; date of current version. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tareq Al-Naffouri. The authors would like to thank the support from the following grants: 973-2015CB351800, NSFC-61231010, NSFC-61527804, NSFC-61210005 and the Microsoft Research Asia Collaborative Research funding.

B. Xin, Y. Wang, and W. Gao are with the Department of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jimxinbo@gmail.com; yizhou.wang@pku.edu.cn; wgao@pku.edu.cn).

D. Wipf is with the Visual Computing group, Microsoft Research, Beijing 100080, China (e-mail: davidwip@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2551697

Before proceeding, we highlight several main contributions as follows:

1) 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. Model justification is ultimately based on such meticulous technical considerations rather than merely the presumed qualitative legitimacy of any underlying prior distributions.
2) Non-convex algorithms have demonstrated some improvement in estimation accuracy over the celebrated convex nuclear norm; however, this typically requires the inclusion of one or more additional tuning parameters to incrementally inject additional objective function curvature and avoid bad local solutions. In contrast, for solving (1) our non-convex Bayesian-inspired algorithm requires no such parameters at all, and noisy relaxations necessitate only a single, standard trade-off parameter balancing data-fit and minimal rank.[1]
3) Over a wide battery of controlled experiments with ground-truth data, our approach outperforms all existing algorithms that we are aware of, Bayesian, non-convex, 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.

II. 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.

A. General Non-Convex Methods

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, [6] 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 $f(z) = \log(z^2+\gamma)$, which maintains an intimate connection with rank given that

$$\log z = \lim_{q\to 0} q^{-1}\left(z^q - 1\right) \quad \text{and} \quad \lim_{q\to 0} z^q = I\left[z \neq 0\right], \quad (4)$$

where I is a standard indicator function. Consequently, when γ is small, $\sum_i \log(\sigma_i[\boldsymbol{X}]^2 + \gamma)$ behaves much like a scaled and translated version of the rank, albeit with nonzero gradients away from zero.

The IRLS0 algorithm from [6] 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, [5] 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 [2]. 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 [5] show that default settings perform relatively poorly.

Finally, a somewhat different class of non-convex algorithms can be derived using a straightforward application of alternating minimization [10]. 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

$$\min_{\boldsymbol{U},\boldsymbol{V}} \| b - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^T\right) \|_{\mathcal{F}} \quad (5)$$

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

[1] While not our emphasis here, similar to other Bayesian frameworks, even this trade-off parameter can ultimately be learned from the data if a true, parameter-free implementation is desired across noise levels.

matrices such that prior knowledge of all individual ranks is not feasible.

B. Bayesian Methods

From a probabilistic perspective, previous work has applied Bayesian formalisms to rank minimization problems, although not specifically within an affine constraint set. For example, [11]–[13] derive robust PCA algorithms built upon the linear summation of a rank penalty and an element-wise sparsity penalty. In particular, [12] applies an MCMC sampling approach for posterior inference, but the resulting iterations are not scalable, subjectable to detailed analysis, nor readily adaptable to affine constraints. In contrast, [11] applies a similar probabilistic model but performs inference using a variational mean-field approximation. While the special case of matrix completion is considered, from an empirical standpoint its estimation accuracy is not competitive with the state-of-the-art non-convex algorithms mentioned above. Finally, without the element-wise sparsity component intrinsic to robust PCA (which is not our focus here), [13] simply collapses to a regular PCA model with a closed-form solution, so the challenges faced in solving (1) do not apply. Consequently, general affine constraints really are a key differentiating factor.

From a motivational angle, the basic probabilistic model with which we begin our development can be interpreted as a carefully re-parameterized generalization of the probabilistic PCA model from [8]. This will ultimately lead to a non-convex algorithm devoid of the heuristic tuning strategies mentioned above, but nonetheless still uniformly superior in terms of estimation accuracy. We emphasize that, although we employ a Bayesian entry point for our algorithmic strategy, final justification of the underlying model will be entirely based on properties of the underlying cost function that emerges, rather than any putative belief in the actual validity of the assumed prior distributions or likelihood function. This is quite unlike the vast majority of existing Bayesian approaches.

C. Analytical Considerations

Turning to analytical issues, a number of celebrated theoretical results dictate conditions whereby substitution of the rank function with the convex nuclear norm in (1) is nonetheless guaranteed to still produce the minimal rank solution. For example, if $\mathcal{A}$ is a Gaussian iid measurement ensemble and $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ represents the optimal solution to (1) with $\mathrm{rank}[\boldsymbol{X}_0] = r$, then with high probability as the problem dimensions grow large, the minimum nuclear norm feasible solution will equal $\boldsymbol{X}_0$ if the number of measurements p satisfies $p \geq 3r(2n-r)$ [14].

The limitation of this type of result is two-fold. First, in the above situation the true minimum rank solution only actually requires $p \geq r(2n-r)$ measurements to be recoverable via brute force solution of (1), and the remaining difference of a factor of three can certainly be considerable in many practical situations (e.g., requiring 300 measurements is far more laborious than only needing 100 measurements). Secondly though, and far more importantly, all existing provable recovery guarantees place extremely strong restrictions on the structure of $\mathcal{A}$, e.g., strong restrictions on the singular value decay of matrices in the nullspace of $\mathcal{A}$. Such conditions are unlikely to ever hold in realistic application settings, including the image rectification example we describe in Section VII.A (in fact, these conditions are usually incapable of even being checked). In contrast, the algorithm we propose is empirically observed to only require the theoretically minimal number of measurements even when such nullspace conditions are violated in many cases. While a general theoretical guarantee of this sort is obviously not possible, we do nonetheless provide several supporting theoretical results indicative of why such performance is at least empirically obtainable.

III. Alternative Algorithm Derivation

In this section we first detail our basic distributional assumptions followed by development of the associated update rules for inference.

A. Basic Model

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

$$p\left(\boldsymbol{b}|\boldsymbol{X};\mathcal{A},\lambda\right) \quad \propto \quad \exp\left[-\frac{1}{2\lambda}\| \mathcal{A}\left(\boldsymbol{X}\right) - \boldsymbol{b} \|_2^2\right], \qquad (6)$$

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

$$p\left(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}\right) = \prod_i \mathcal{N}\left(\boldsymbol{x}_{:i};\boldsymbol{0},\nu_i\boldsymbol{\Psi}\right) \propto \exp\left[\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x}\right], \quad (7)$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is a positive semi-definite symmetric matrix,[2] $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_m]^\top$ is a non-negative vector, $\boldsymbol{x} = \mathrm{vec}[\boldsymbol{X}]$ (column-wise vectorization), and $\overline{\boldsymbol{\Psi}} = \mathrm{diag}[\boldsymbol{\nu}] \otimes \boldsymbol{\Psi}$, 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 (7) 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 $\widehat{\boldsymbol{X}}$ such that

$$\widehat{\boldsymbol{x}} = \mathrm{vec}\left[\widehat{\boldsymbol{X}}\right] = \overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\left(\lambda\boldsymbol{I} + \boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{b}. \qquad (8)$$

[2]Technically $\boldsymbol{\Psi}$ must be positive definite for the inverse in (7) to be defined. However, we can accommodate the semi-definite case using the following convention. Without loss of generality assume that $\overline{\boldsymbol{\Psi}} = \boldsymbol{R}\boldsymbol{R}^\top$ for some matrix $\boldsymbol{R}$. We then qualify that $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) = 0$ if $\boldsymbol{x} \notin \mathrm{span}[\boldsymbol{R}]$, and $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) \propto \exp[-\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}]$ otherwise. Equivalently, throughout the paper for convenience (and with slight abuse of notation) we define $\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \infty$ when $\boldsymbol{x} \notin \mathrm{span}[\boldsymbol{R}]$, and $\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}$ otherwise. This will come in handy, for example, when interpreting the bound in (12) below. Note also that the final cost function (10) we will ultimately be minimizing requires no such inverse anyway.

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 $\widehat{\boldsymbol{X}}$ 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 $\boldsymbol{\nu}$ [15], [13], [16]. This involves solving

$$\max_{\boldsymbol{\Psi} \in H^+, \boldsymbol{\nu} \geq 0} \int p\left(\boldsymbol{b}|\boldsymbol{X}; \mathcal{A}, \lambda\right) p\left(\boldsymbol{X}; \boldsymbol{\Psi}, \boldsymbol{\nu}\right) d\boldsymbol{X}, \tag{9}$$

where H^+ denotes the set of positive semi-definite and symmetric $n \times n$ matrices. After a $-2\log$ transformation and application of a standard convolution-of-Gaussians integration, solving (9) is equivalent to minimizing the cost function

$$\mathcal{L}\left(\boldsymbol{\Psi}, \boldsymbol{\nu}\right) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} + \log|\boldsymbol{\Sigma}_b|, \tag{10}$$

where

$$\boldsymbol{\Sigma}_b = \boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda \boldsymbol{I} \text{ and } \overline{\boldsymbol{\Psi}} = \text{diag}\left[\boldsymbol{\nu}\right] \otimes \boldsymbol{\Psi}. \tag{11}$$

Here $\boldsymbol{\Sigma}_b$ is the covariance of $\boldsymbol{b}$ given $\boldsymbol{\Psi}$ and $\boldsymbol{\nu}$.

B. Update Rules

Minimizing (10) is a non-convex optimization problem, and we employ standard upper bounds for this purpose leading to an EM-like algorithm, somewhat related to [8]. 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 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 [16], for the first term in (10) we have

$$\boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \boldsymbol{x}^\top \overline{\boldsymbol{\Psi}}^{-1} \boldsymbol{x} \tag{12}$$

with equality whenever $\boldsymbol{x}$ satisfies (8). For the second term we use

$$\begin{aligned} \log|\boldsymbol{\Sigma}_b| &\equiv m \log|\boldsymbol{\Psi}| + \log|\lambda \boldsymbol{A}^\top \boldsymbol{A} + \overline{\boldsymbol{\Psi}}^{-1}| \\ &\leq m \log|\boldsymbol{\Psi}| + \text{tr}\left[\boldsymbol{\Psi}^{-1} \nabla_{\Psi^{-1}}\right] + C, \end{aligned} \tag{13}$$

where because $\log|\lambda \boldsymbol{A}^\top \boldsymbol{A} + \overline{\boldsymbol{\Psi}}^{-1}|$ 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

$$\nabla_{\Psi^{-1}} = \sum_{i=1}^{m} \boldsymbol{\Psi} - \boldsymbol{\Psi}\boldsymbol{A}_i^\top \left(\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_i \boldsymbol{\Psi}, \tag{14}$$

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 (12) and (13) 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

$$\begin{aligned} \boldsymbol{\Psi}^{\text{opt}} &= \arg\min_{\boldsymbol{X}} \text{tr}\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{X}\boldsymbol{X}^\top + \nabla_{\Psi^{-1}}\right)\right] + m \log|\boldsymbol{\Psi}| \\ &= \frac{1}{m}\left[\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top + \nabla_{\Psi^{-1}}\right]. \end{aligned} \tag{15}$$

By agnostically starting with $\boldsymbol{\Psi} = \boldsymbol{I}$ and then iteratively computing (8), (14), and (15), we can then obtain an estimate for $\boldsymbol{\Psi}$, and more importantly, a corresponding estimate for $\boldsymbol{X}$ given by (8) 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).

IV. Properties of BARM

Here we first describe a close but perhaps not intuitively-obvious relationship between the BARM objective function and canonical nuclear norm minimization. We then discuss desirable properties of global and local minima before concluding with a brief examination of convergence issues.

A. Connections with Nuclear Norm Minimization

On the surface, it may appear that minimizing (10) is completely unrelated to the convex problem

$$\min_{\boldsymbol{X}} \| \boldsymbol{X} \|_* \text{ s.t. } \boldsymbol{b} = \mathcal{A}\left(\boldsymbol{X}\right) \tag{16}$$

that is most commonly associated with practical rank minimization implementations. However, a close connection can be revealed by considering the modified objective function

$$\mathcal{L}'\left(\boldsymbol{\Psi}\right) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} + \text{tr}\left[\overline{\boldsymbol{\Psi}}\right], \tag{17}$$

which represents nothing more than (10), with $\boldsymbol{\nu} = \boldsymbol{1}$ and with $\log|\boldsymbol{\Sigma}_b|$ being replaced by $\text{tr}[\overline{\boldsymbol{\Psi}}]$. Now suppose we minimize (17) with respect to $\boldsymbol{\Psi} \in H^+$ obtaining some $\boldsymbol{\Psi}^*$. We then go on to compute an estimate of $\boldsymbol{X}$ using (8). Note that if we apply the bound from (12) to the first term in (17), then this estimate for $\boldsymbol{X}$ equivalently solves

$$\min_{\boldsymbol{\Psi} \in H^+, \boldsymbol{X}} \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \boldsymbol{x}^\top \overline{\boldsymbol{\Psi}}^{-1} \boldsymbol{x} + \text{tr}\left[\overline{\boldsymbol{\Psi}}\right], \tag{18}$$

with $\boldsymbol{x} = \text{vec}[\boldsymbol{X}]$ as before. If we first optimize over $\boldsymbol{\Psi}$, it is easily demonstrated that the optimal value of $\boldsymbol{\Psi}$ equals $\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{1/2}$. Plugging this value into (18), simplifying, and then applying the definition of the nuclear norm, we arrive at

$$\min_{\boldsymbol{X}} \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + 2\| \boldsymbol{X} \|_*, \tag{19}$$

Furthermore, in the limit $\lambda \to 0$ (applied outside of the minimization), (19) becomes equivalent to (16). For more information regarding the duality relationship between variance/covariance space and coefficient space, at least in the related context of compressive sensing models, please refer to [16].

Consequently, we may conclude that the central distinction between the proposed BARM cost function and nuclear norm minimization is an intrinsic $\mathcal{A}$-dependent penalty function

$\log|\boldsymbol{\Sigma}_b|$ which is applied in covariance space. In Section IV.B we will examine desirable properties of this non-convex substitution, highlighting our desire to treat the underlying BARM probabilistic model as an independent cost function that may be subject to technical analysis independent of its Bayesian origins.

B. Global/Local Minima Analysis

As discussed in Section II one nice property of the $\sum_i \log(\sigma_i[\boldsymbol{X}])$ penalty employed (approximately) by IRLS0 [6] 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 $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ 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 $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ with $\gamma \to 0$. Before presenting these results, we define $\mathrm{spark}[\boldsymbol{A}]$ as the smallest number of linearly dependent columns in matrix $\boldsymbol{A}$ [17]. All proofs are deferred to the Appendix.

Lemma 1: Let $\boldsymbol{b}=\boldsymbol{A}\mathrm{vec}[\boldsymbol{X}]$, where $\boldsymbol{A}\in\mathbb{R}^{p\times \mathrm{nm}}$ 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 (10) in the limit $\lambda\to 0$ is such that $\boldsymbol{x}^*=\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top(\boldsymbol{A}\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top)^\dagger\boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*]=r$ with $\mathrm{vec}[\boldsymbol{X}^*]=\boldsymbol{x}^*$.

Lemma 2: Additionally, let $\widetilde{\boldsymbol{A}}=\boldsymbol{A}D$, 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 (10) in the limit $\lambda\to 0$, then $\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}^*,\mathrm{diag}[\boldsymbol{\alpha}]^{-1}\boldsymbol{\nu}^*\}$ is a minimizer when $\widetilde{\boldsymbol{A}}$ 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 $p>\mathrm{rm}$. 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 II).

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 (9), 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 after some transformation $\boldsymbol{W}$, 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 $\cap_i\mathrm{null}[\boldsymbol{A}_i]=\varnothing$. Then if $\min_{\boldsymbol{X}}\mathrm{rank}[\boldsymbol{X}]=1$ in the feasible region, any minimizer $\{\boldsymbol{\Psi}^*,\boldsymbol{\nu}^*\}$ of (10) (global or local) in the limit $\lambda\to 0$ is such that $\boldsymbol{x}^*=\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top(\boldsymbol{A}\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top)^\dagger\boldsymbol{b}$ 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, at least in this rank-one case. (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.[3]

Although a global optimal guarantee for finding a rank-one matrix sounds somewhat limited, 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. Moreover, finding rank one matrices subject to affine constraints represents a crucial component of applications such as phase retrieval [18], [19].

Additionally, if a unique rank-one solution exists to (1), then the unique minimizing solution to (10) will produce this $\boldsymbol{X}$ via (8). 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.[4] Moreover, as evident from the experiments, the proposed algorithm always successfully finds the global optimal in many situations where the underlying matrix has a rank much higher than one. Therefore, although we can only provide theoretical guarantee for the rank-one case, the underlying intuition that local minima are smoothed away arguably carries over to situations where the rank is greater than one.

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

[4]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.

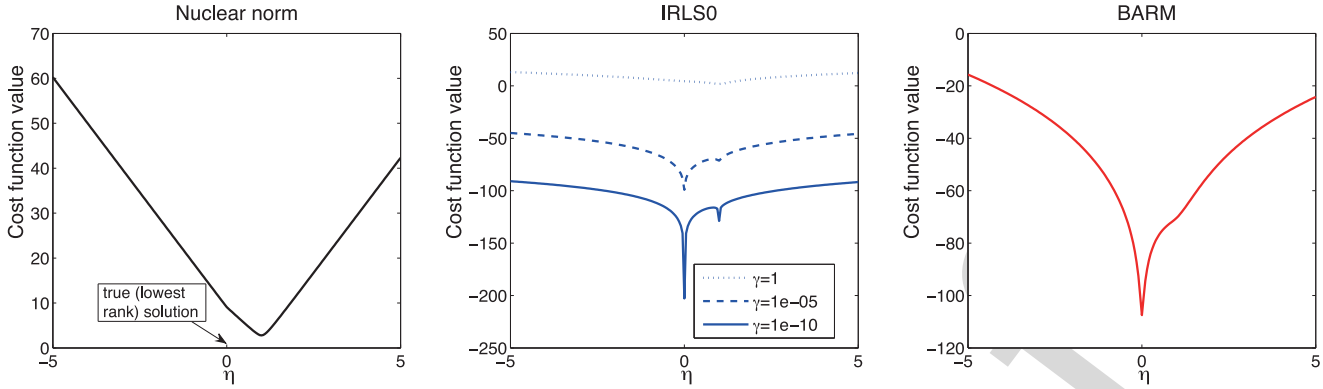


Fig. 1. Plots of different surrogates for matrix rank in a 1D feasible subspace. Here the convex nuclear norm does not retain the correct global minimum. In contrast, although the non-convex $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ penalty exhibits the correct minimum when γ is sufficiently small, it also contains spurious minima. Only BARM smoothes away local minimum while simultaneously retaining the correct global optima.

C. Visualization of BARM Local Minima Smoothing

To further explore the smoothing effect and complement Theorem 1, it helps to visualize rank penalty functions restricted to the feasible region. While the BARM algorithm involves minimizing (10), its implicit penalty function on $\boldsymbol{X}$ can nonetheless be numerically obtained across the feasible region in a given subspace of interest; for other penalties such as the nuclear norm this is of course trivial. Practically it is convenient to explore a 1D feasible subspace generated by $\boldsymbol{X}^*+\eta\boldsymbol{V}$, where $\boldsymbol{X}^*$ is the true minimum rank solution, $\boldsymbol{V}\in\text{null}[\boldsymbol{A}]$, and η is a scalar. We may then plot various penalty function values as η is varied, tracing the corresponding 1D feasible subspace. We choose $\boldsymbol{V}=\boldsymbol{X}^1-\boldsymbol{X}^*$, where $\boldsymbol{X}^1$ is a feasible solution with minimum nuclear norm; however, random selections from $\text{null}[\boldsymbol{A}]$ also show similar characteristics.

Fig. 1 provides a simple example of this process. $\boldsymbol{A}$ is generated randomly with all zeros and a single randomly placed '1' in each row leading to a canonical matrix completion problem. $\boldsymbol{X}^*\in\mathbb{R}^{5\times 5}$ is randomly generated as $\boldsymbol{X}^*=\boldsymbol{u}\boldsymbol{v}^\top$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are iid $\mathcal{N}(0,1)$ vectors, and so $\boldsymbol{X}^*$ is rank one. Finally, $p=10$ elements are observed, and therefore $\boldsymbol{A}$ has 10 rows and $5\times 5=25$ columns. η is varied from -5 to 5 and the values of the nuclear norm, $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$, and the implicit BARM cost function are displayed.

From the figure we observe that the minimum of the nuclear norm is not produced when the rank is smallest, which occurs when $\eta=0$; hence the convex cost function fails for this problem. Likewise, the $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ penalty used by IRLS0 displays an incorrect global minimum when the tuning parameter γ is large. In contrast, when γ is small, while the global minimum may now be correct, spurious local ditches have appeared in the cost function.[5] Therefore, any success of the IRLS0 algorithm depends heavily on a carefully balanced decaying sequence of γ values, with the hope that initial iterations can steer the trajectory towards a desirable basin of attraction where local minima are less problematic. One advantage of BARM then is that it is parameter free in this respect and yet still retains the correct global minimum, often without additional spurious local minima.

[5]Technically speaking, these are not provably local minima since we are only considering a 1D subspace of the feasible region. However, it nonetheless illustrates the strong potential for troublesome local minima, especially in high dimensional practical problems.

D. Convergence

Previous results of Section IV are limited to exploring aspects of the underlying BARM cost function. Regarding the BARM algorithm itself, by construction the updates generated by (8), (14), and (15) 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 [20]. However, provided we add a small regularization factor $\gamma\text{tr}[\boldsymbol{\Psi}^{-1}]$, 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 $\boldsymbol{\Psi}^*\in\mathcal{S}$, $\mathcal{L}(\boldsymbol{\Psi}^*)+\gamma\text{tr}[(\boldsymbol{\Psi}^*)^{-1}]$ 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 [6] as discussed in Section II. 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.

V. 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 $\boldsymbol{x} = \text{vec}[\boldsymbol{X}]$ 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

$$\overline{\boldsymbol{\Psi}} = 1/2\left(\boldsymbol{\Psi}_r \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{\Psi}_c\right). \tag{20}$$

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 listed in the Appendix and computational complexity evaluated in Section VI.E.

VI. Experimental Validation

This section compares BARM with existing state-of-the-art affine rank minimization algorithms. For BARM, in all noiseless cases we simply used $\lambda = 10^{-10}$ (effectively zero), and hence no tuning parameters are required. Likewise, nuclear norm minimization [1], [4] 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 [6]. 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 [5] where available. Note that both [5] and [6] 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 [10] or the normalized hard thresholding algorithm from [21]. 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. However, we do show limited empirical results with the variational sparse Bayesian algorithm (VSBL) from [11] 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.

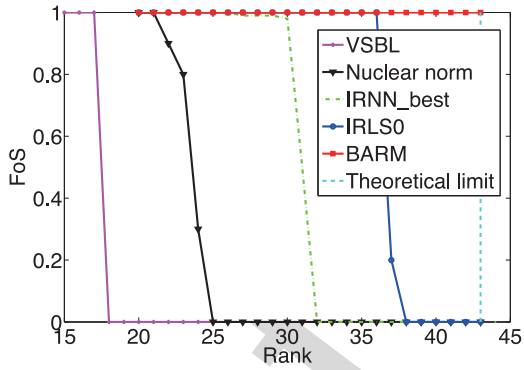


Fig. 2. Matrix completion comparisons (avg of 10 trials).

A. 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 [5], where a rank r matrix is generated as $\boldsymbol{X}_0 = \mathbf{TM}_L\mathbf{TM}_R$, with $\mathbf{TM}_L \in \mathbb{R}^{n\times r}$ and $\mathbf{TM}_R \in \mathbb{R}^{r\times m}$ $(n = m = 150)$ as iid $\mathcal{N}(0,1)$ random matrices. 50% of all entries are then hidden uniformly at random. The *relative error* (REL) given by $\| \boldsymbol{X}_0 - \widehat{\boldsymbol{X}}\|_{\mathcal{F}}/\| \boldsymbol{X}_0 \|_{\mathcal{F}}$ 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 Fig. 2 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 [5] 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 [6] 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 [6] tests a variety of

TABLE I
MATRIX COMPLETION RESULTS OF BARM WITH IRLS0 ON THE THREE HARDEST PROBLEMS FROM [6]. PUBLISHED RESULTS IN [6] 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

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 [6]. We also only include experiments where algorithms are blind to the true rank of $\boldsymbol{X}_0$.[6] Results are shown in Table I, where we have also displayed the published results of three additional algorithms that were compared with IRLS0 in [6], namely, IHT [22], FPCA [23] and Optspace [24]. From the table we observe that, in the most difficult problem considered in [6], IRLS0 achieved only a 0.5 FoS score (meaning failure 50% of the time) while BARM still achieves a perfect 1.0. Note that when FR is high, the problem of recovering the underlying matrix is essentially much harder. This happens in a manner that more local minima are induced (due to increased rank) and/or much larger search space are exposed (due to decreased number of observations/constraints). In these cases, the equivalency of the global optimal with convex relaxation usually does not hold, whereas for the existing non-convex surrogates, there is no reason to assume any local minima are not present. However, since BARM has an implicit mechanism of smoothing local minima (though maybe not all of them), it works more robustly in these situations.

B. General $\boldsymbol{A}$

Next we consider the more challenging problem involving arbitrary affine constraints. The desired low-rank $\mathbf{T}\mathbf{X}_0$ 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. Fig. 3 displays aggregate results when $\boldsymbol{X}_0$ is 50×50 and 100×100, including the underlying REL scores for additional comparison. 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.

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

C. Failure Case Analysis

Thus far we have not shown any cases where BARM actually fails. Of course solving (1) for general $\boldsymbol{A}$ 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 $p = 975$.

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. In Fig. 4 we have averaged the singular values of $\widehat{\boldsymbol{X}}$ in all the 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. 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 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 $\widehat{\boldsymbol{X}}$ is found such that $\sigma_r[\widehat{\boldsymbol{X}}]/\sigma_{r+1}[\widehat{\boldsymbol{X}}] > 10^3$, 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

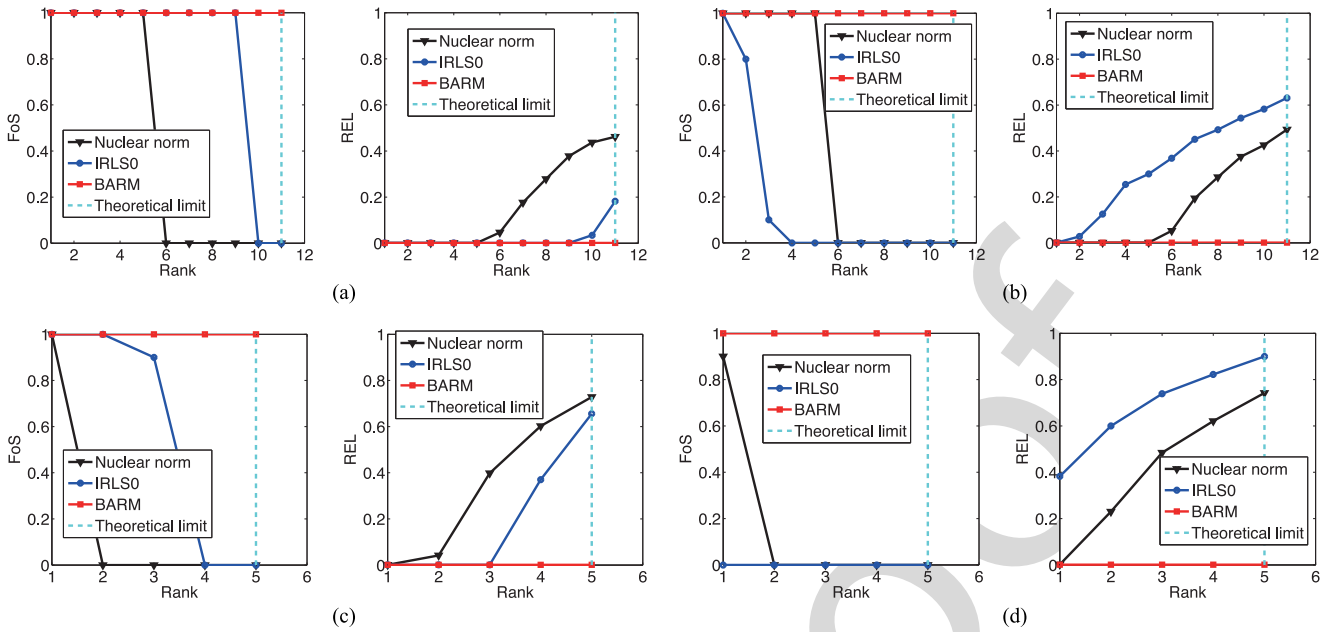


Fig. 3. Comparisons with general affine constraints (avg of 10 trials). (a) 50×50, $\boldsymbol{A}$ uncorrelated, (b) 50×50, $\boldsymbol{A}$ correlated, (c) 100×100, $\boldsymbol{A}$ uncorrelated, and (d)100×100, $\boldsymbol{A}$ correlated.

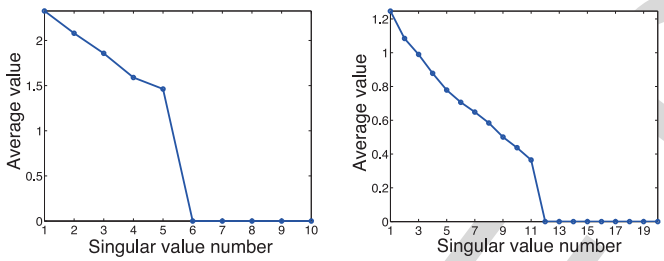


Fig. 4. Singular value averages of failure cases. In both cases solutions of minimal rank are obtained even though $\widehat{\boldsymbol{X}} \neq \boldsymbol{X}_0$. (a) 50×50 and (b) 100×100.

TABLE II
FURTHER MATRIX COMPLETION COMPARISONS OF BARM WITH IRLS0 BY REDUCING THE NUMBER OF MEASUREMENTS IN THE HARDEST PROBLEM FROM [6]. 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

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 I the most difficult case involves $\mathrm{FR} = 0.87$; 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 II. While IRLS0 (which is the top performing algorithm in [6] and in our experiments besides BARM) fails 100% of the time via both metrics, BARM can achieve an FoS of 0.7 even when $\mathrm{FR} = 0.99$ 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. Fig. 5 depicts the results using problem sizes consistent with Fig. 3; note that the FoRS metric has replaced the REL metric for comparison purposes.

Two things stand out from the 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 $\widehat{\boldsymbol{X}}$ 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

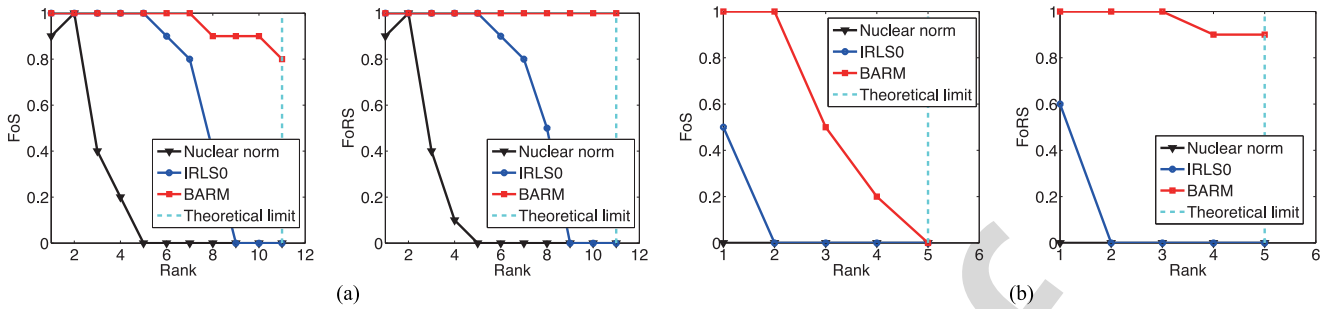


Fig. 5. Comparisons with structured affine constraints using both FoS and FoFS evaluation metrics (avg of 10 trials). (a) 50×50, $\boldsymbol{A}$ sub-sampled DCT, (b) 100×100, $\boldsymbol{A}$ sub-sampled DCT.

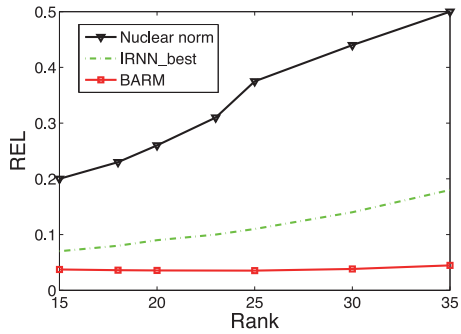


Fig. 6. Test with noisy data.

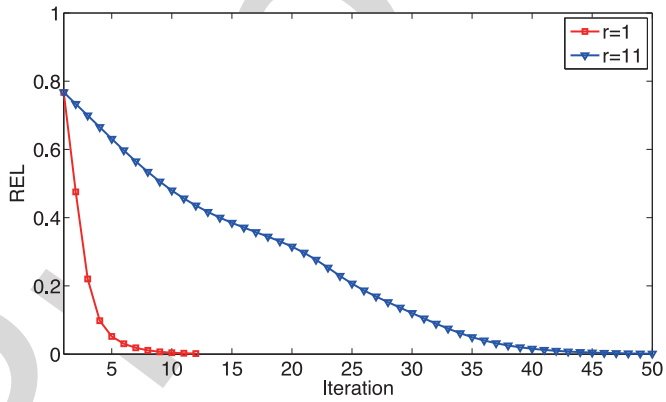


Fig. 7. Empirical convergence of BARM.

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.

D. Additional Noisy Tests

We also briefly present results that demonstrate the robustness of BARM to noise. For this purpose we reproduce the noisy experiment from [5] designed for validating IRNN algorithms. The simulated data are generated in the exact same way as was used to produce Fig. 2, only now instead of observing elements of $\boldsymbol{X}_0$ directly, we observe $\boldsymbol{X}_0 + 0.1 \times \boldsymbol{E}$, where elements of $\boldsymbol{E}$ are iid $\mathcal{N}(0,1)$. Although in [5] a heuristic strategy is introduced and tuned for adaptively setting all parameters (four in total), we simply applied BARM with $\lambda = 10^{-3}$ (so only a single parameter need be adjusted, and actually a wide range of λ values produces similar performance anyway). Results are shown in Fig. 6 where we compare BARM directly with the best result reported in [5] over the range $r = 15$ to $r = 35$. The nuclear norm solution is also included for reference. Overall, the BARM solution is stable and exhibits superior accuracy relative to the others.

E. Computational Complexity

Finally, regarding computational complexity, for general $\boldsymbol{A}$ the BARM updates can be implemented to scale linearly in the elements of $\boldsymbol{X}$ and quadratically in the number of observations p (the special case of matrix completion is decidedly much cheaper because of the special structure that can be exploited). In our experiments, for relatively easy problems on the order of 10 iterations are required, while for difficult recovery problems near the theoretical recovery boundary this may increase by a factor of 10 or so. This is somewhat expected though since as we near the theoretical limit, $\boldsymbol{A}$ becomes highly overcomplete, and candidate solutions become much more difficult to differentiate.

To show this effect empirically, we compare two separate trials from Fig. 3(a), the first when $r = 1$ (relatively easy), the second when $r = 11$ (relatively hard).[7] In Fig. 7 we plot the value of REL in both cases versus the iteration number of BARM.

VII. Application Examples

Many real-world problems from disparate fields can be formulated as the search for a low-rank matrix under affine constraints [1], [3], [4], [25]. Here we briefly consider two such examples: low-rank 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.

A. Low-Rank Image Rectification

In [4], the *transform invariant low-rank textures* (TILT) algorithm is derived for rectifying images containing low-rank

[7] Note that $r = 1$ is only relatively easy here because the number of observations is sufficient for the larger $r = 11$ case; if only the minimal number of measurements are available then even $r = 1$ can be challenging for many algorithms.

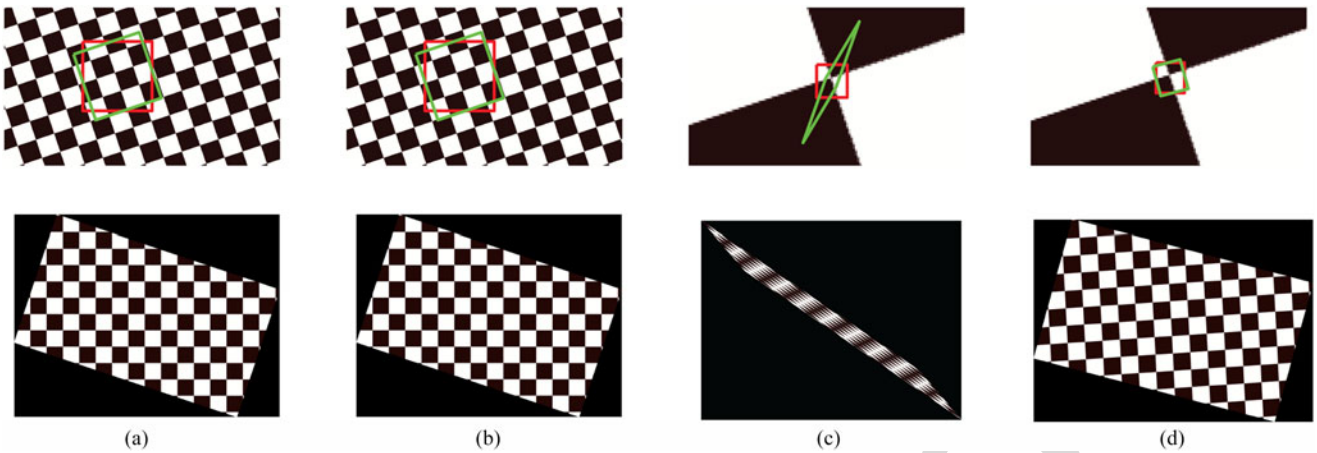


Fig. 8. Image rectification comparisons using a checkboard image. *Top*: Original image with observed region (red box) and estimated transformation (green box). *Bottom*: Rectified image estimates. (a) Nuclear norm (easy), (b) BARM (easy), (c) Nuclear norm (hard), (d) BARM (hard).

textures that have been transformed using an unknown operator $\boldsymbol{\tau}$ from some group (e.g., a homography). For a given observed image $\boldsymbol{Y}$, the basic idea is to construct a first-order Taylor series approximation around the current rectified image estimate $\widehat{\boldsymbol{X}}$ and solve

$$\min_{\boldsymbol{X},\boldsymbol{\delta}} \operatorname{rank}[\boldsymbol{X}] \text{ s.t. } \boldsymbol{X} = \boldsymbol{Y} + \sum_i \boldsymbol{J}_i\left(\widehat{\boldsymbol{X}}\right)\delta_i, \tag{21}$$

where $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$ is the Jacobian matrix with respect to $\boldsymbol{X}$ of the i-th parameter τ_i describing the transformation, with $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots]^\top$. Optimization over the vector of first-order differences $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots]^\top$ can be accomplished in closed form by projecting both sides of the constraint to the orthogonal complement of the span of all $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$. Let $P_{\boldsymbol{J}^c}$ represent this projection operator. The feasible region in (21) then becomes

$$P_{\boldsymbol{J}^c}(\boldsymbol{X}) = P_{\boldsymbol{J}^c}(\boldsymbol{Y}) + P_{\boldsymbol{J}^c}\left(\sum_i \boldsymbol{J}_i\left(\widehat{\boldsymbol{X}}\right)\delta_i\right) = P_{\boldsymbol{J}^c}(\boldsymbol{Y}) \tag{22}$$

The resulting problem then reduces exactly to (1) when we define $\mathcal{A} = P_{\boldsymbol{J}^c}$ and $\boldsymbol{b} = \operatorname{vec}[P_{\boldsymbol{J}^c}(\boldsymbol{Y})]$. Once $\boldsymbol{X}$ is computed in this way, we then update each $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$ and repeat until convergence.

While the original TILT algorithm substitutes the nuclear norm for $\operatorname{rank}[\boldsymbol{X}]$, we embedded the BARM algorithm into the posted TILT source code [4] for comparison purposes (note that we disabled an additional sparse error term for both algorithms to simplify comparisons, and it is not necessary anyway in many regimes). Figs. 8 and 9 display results on both two easy examples, where the number of observations p is large, and two more difficult problems where the number observations is small. While both algorithms succeed on the easy cases, when the observations are constrained by a small image window, only BARM is successful in accurately rectifying the images. This may be due, at least in part, to the fact that the implicit $\mathcal{A}$ operator contains significant structure that is not consistent with the required nullspace properties required for nuclear norm minimization success.

B. Collaborative Filtering of MovieLens Data

Collaborative filtering, a technique used by many recommender systems, is a popular representative application of low-rank matrix completion. Typically the rows (or columns) of $\boldsymbol{X}_0$ index users, the columns (or rows) denote items, and each entry $(\boldsymbol{X}_0)_{\text{ij}}$ is the rating/score of user i applied to item j. Given that we can observe some subset of elements of $\boldsymbol{X}_0$, the task of collaborative filtering is to predict all or some of the missing ratings. In general this would be impossible; however, if we have access to some prior knowledge, e.g., $\boldsymbol{X}_0$ is low-rank, then estimation may be feasible.

While our interest here is not in recommender systems or collaborative filtering per se, we nonetheless evaluate BARM using the 1M MovieLens dataset[8] as this appears to represent one of the most common evaluation benchmarks. We emphasize at the outset that the strict validity of any low-rank assumptions underlying this data is debatable, and it remains entirely unclear whether the true globally optimal or lowest rank solution consistent with the observations, even if computable, would necessarily lead to the best prediction of the unknown ratings. In fact, the reported performance of various existing rank-minimization algorithms tends to cluster around almost the same value, implying that collaborative filtering may not provide the most discriminative data type with which to compare. In most cases, it appears that tuning parameters and other heuristic modifications play a larger role than the underlying algorithmic distinctions fundamental to finding optimal low-rank estimates. Nonetheless, we apply BARM for completeness and convention, adopting an additional simple mean-offset estimation term from [25] that is particularly suitable for this problem.

In [6], IRLS0 is compared with only two other algorithms on MovieLens data, but the performance is no better. Therefore, we choose to compare directly with [25], which both derives an IRLS-like algorithm and shows comparisons with a much wider variety of alternative algorithms using a strict evaluation protocol that is standard in the literature. Specifically, the

[8]http://www.grouplens.org/

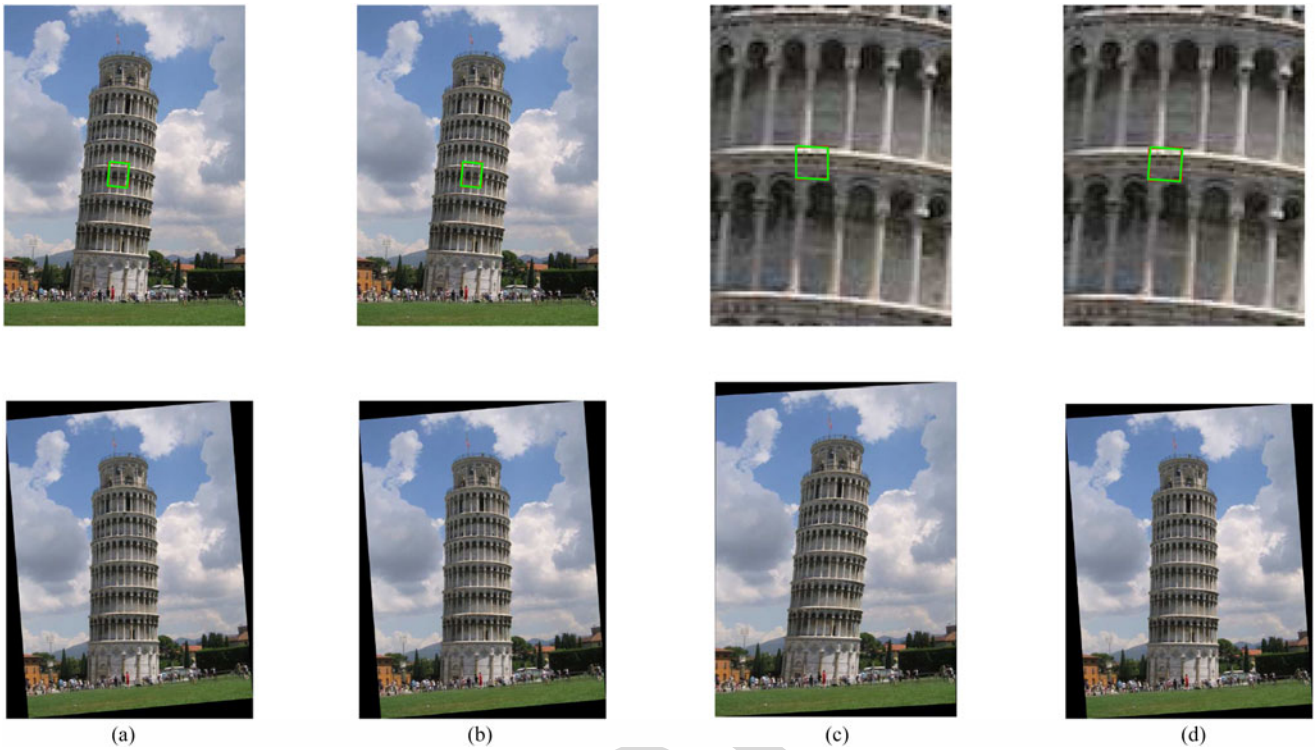


Fig. 9. Image rectification comparisons using a landmark photo. *Top*: Original image with observed region (red box) and estimated transformation (green box). *Bottom*: Rectified image estimates. (a) Nuclear norm (easy), (b) BARM (easy), (c) Nuclear norm (hard), (d) BARM (hard).

1M MovieLens dataset, which contains 1 million ratings in the range $\{1, \ldots, 5\}$ for 3900 movies from 6040 unique users, is assessed under two test-protocals: *weak generalization*, which measures the ability to predict other items rated by the same user, and *strong generalization*, which measures the ability to predict items by novel users. 5 000 users are randomly selected for the weak generalization, and likewise 1 000 users are extracted for the strong generalization. Each experiment is then run three times and the averaged results are reported. The performance metric is *normalized mean absolute error* (NMAE) given as

$$\text{NMAE} = \frac{\left(\sum_{i,j \in \text{supp}(\boldsymbol{X}_0)} \frac{|(\boldsymbol{X}_0)_{\text{ij}} - \widehat{\boldsymbol{X}}_{\text{ij}}|}{|\text{supp}(\boldsymbol{X}_0)|} \right)}{(rt_{\max} - rt_{\min})},$$

where $rt_{\max}$ and $rt_{\min}$ are the maximum and minimum ratings possible.

We followed the same setup and reported results using BARM in Table III along with results from [25] for comparison. This includes the additional algorithms URP [26], Attitude [27], MMMF [28], IPCF [29], E-MMMF [30], GPLVM [31], NBMC [32], and IRLS/GM [25], [6]. From this table we observe that for the easier weak generalization problem BARM is a close second best, while for the more challenging strong generalization BARM is actually the best. Of course it is also immediately apparent that all algorithms fall within a relatively narrow performance range of approximately five percentage points. Consequently, we cannot unequivocally conclude that the attributes of BARM which make it suitable for optimally minimizing rank

TABLE III
COLLABORATIVE FILTERING ON 1M MOVIELENS DATASET. RESULTS FROM [25] ARE IN ITALIC FOR COMPARISON PURPOSES

	Weak NMAE	Hard NMAE
URP	*0.4341*	*0.4444*
Attitude	*0.4320*	*0.4375*
MMMF	*0.4156*	*0.4203*
IPCF	*0.4096*	*0.4113*
E-MMMF	*0.4029*	*0.4071*
GPLVM	*0.4026*	*0.3994*
NBMC	**0.3916**	*0.3992*
IRLS/GM	*0.3959*	*0.3928*
BARM	0.3942	**0.3898**

necessarily translate into a truly significant practical advantage on this collaborative filtering task. But we would argue that the same holds for any matrix completion algorithm.

VIII. CONCLUSION

This paper explores a conceptually-simple, parameter-free algorithm called BARM 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. Our strategy in this effort has been to adopt Bayesian machinery for inspiring a principled cost function; however, ultimate model justification is placed entirely in

theoretical evaluation of desirable global and local minima properties, and in the empirical recovery performance that inevitably results from these properties. Although in general non-convex algorithms are exponentially more challenging to analyze, in this regard we have at least attempted to contextualize BARM in the same manner as convex optimization-based approaches such as nuclear-norm minimization.

APPENDIX A

Here we provide brief proofs of Lemmas 1 and 2 as well as Theorem 1. We also address the augmented update rules that account for the revised, symmetrized cost function discussed in Section V.

A. Proof of Lemmas 1 and 2

Regarding Lemma 1, this result mirrors related ideas from [16] in the context of Bayesian compressive sensing. Hence, while a more rigorous presentation is possible, here we describe the basic aspects of the adaptation. At any candidate minimizer of (10) in the limit $\lambda \to 0$, define $\boldsymbol{W}$ such that $\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top = \boldsymbol{W}\boldsymbol{W}^\top$. To be a minimizer, global or local, it must be that $\boldsymbol{b} \in \text{span}[\boldsymbol{W}]$. If this were not the case, then $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ would diverge to infinity as $\lambda \to 0$ because $\boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}$ progresses to infinity at a faster rate than $\log |\boldsymbol{\Sigma}_b|$ can compensate by approaching minus infinity. Intuitively, in much the same way $\text{argmin}_z \frac{1}{z} + \log z = 1$, meaning the optimal z must lie in the 'span' of 1 else the overall objective will be driven to infinity.

Consequently, the only way to minimize the cost in the limit as $\lambda \to 0$ is to consider low-rank solutions within the constraint set that $\boldsymbol{b} \in \text{span}[\boldsymbol{W}]$, and it is equivalent to requiring that $\boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq C$ for some constant C independent of λ (which ultimately corresponds with maintaining $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ in the limit as well).

In this setting, while $0 \leq \boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq C$ is bounded, the second term in $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ can be unbounded from below when $\text{rank}[\boldsymbol{\Psi}]$ is sufficiently small. To see this note that

$$\log |\boldsymbol{\Sigma}_b| = \sum_{i=1}^{p} \log \left(\sigma_i \left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top \right] + \lambda \right), \tag{23}$$

where $\sigma_i[\cdot]$ denotes the i-th singular value of a matrix. While the maximum rank of $\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top$ is obviously p, if $r \triangleq \text{rank}[\boldsymbol{\Psi}] < p/m$ and $\text{spark}[\boldsymbol{A}] = p+1$ (maximal spark) as stipulated in the lemma statement, then $\text{rank}\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] = mr$ and (23) becomes

$$\log |\boldsymbol{\Sigma}_b| = \sum_{i=1}^{mr} \log \left(\sigma_i \left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top \right] + \lambda \right) + (p - mr) \log \lambda. \tag{24}$$

Note that the spark assumption accomplishes two objectives in this context. First, it guarantees that a high rank $\boldsymbol{\Psi}$ cannot masquerade as a low rank $\boldsymbol{\Psi}$ behind the nullspace of some collection of columns $\boldsymbol{A}_i$. Secondly, it ensures that after assuming $r < p/m$, then $\text{rank}\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] = mr$.

Consequently, in the limit where $\lambda \to 0$ (with the limit being taken outside of the minimization), (23) effectively scales as $(p - mr) \log \lambda$, and hence the overall cost is minimized when $\boldsymbol{\Psi}$ has minimal rank. This in turn ensures that the corresponding $\boldsymbol{X}$ will also have minimal rank, completing the proof sketch for Lemma 1.

Finally, Lemma 2 follows directly from the structure of the $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ cost function via simple reparameterizations. ■

B. Proof of Theorem 1

To begin we assume that $\boldsymbol{b}_i \neq 0$, $\forall i$, where $\boldsymbol{b}_i$ denotes the sub-vector of $\boldsymbol{b}$ such that $\boldsymbol{b}_i = \boldsymbol{A}_i \boldsymbol{x}_{:i}$. If this were not the case we can always collapse $\boldsymbol{X}$ by the corresponding column (which is indistinguishable from zero) and achieve an equivalent result. Given the assumptions of Theorem 1, the BARM cost function becomes

$$\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu}) = \sum_{i=1}^{m} \boldsymbol{b}_i^\top \left(\nu_i \boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top \right)^{-1} \boldsymbol{b}_i + \log \left| \nu_i \boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top \right|. \tag{25}$$

If there exists a feasible rank one solution to $\boldsymbol{b} = \boldsymbol{A}\text{vec}[\boldsymbol{X}]$, then there also exists a set of $\boldsymbol{\Psi}_i' = \nu_i \boldsymbol{\Psi}$ such that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \boldsymbol{A}_i \boldsymbol{\Psi}_i' \boldsymbol{A}_i^\top$ for all i. To see this, note that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \boldsymbol{A}_i \boldsymbol{x}_{:i} \boldsymbol{x}_{:i}^\top \boldsymbol{A}_i^\top$. Because $\text{rank}[\boldsymbol{X}] = 1$, it also follows that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \alpha_i \boldsymbol{A}_i \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_i^\top$, where $\alpha_i = \| \boldsymbol{x}_{:i} \boldsymbol{x}_{:i}^\top \| / \| \boldsymbol{X}\boldsymbol{X}^\top \|$. Therefore $\boldsymbol{\Psi}_i' = \nu_i \boldsymbol{X}\boldsymbol{X}^\top$ achieves the desired result with $\nu_i = \alpha_i$.

Now suppose we have converged to any solution $\{\widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\nu}}\}$ with $\text{rank}[\boldsymbol{\Psi}] > 1$ and associated $\widehat{\overline{\boldsymbol{\Psi}}} = \boldsymbol{I} \otimes \widehat{\boldsymbol{\Psi}}$. Note that since $\boldsymbol{b}_i \neq 0$, $\boldsymbol{\nu}_i > 0$ for all i, otherwise a local minimum is not possible (the cost function would be driven to positive infinity).

Define $\widehat{\boldsymbol{\Sigma}}_{b_i} = \hat{\nu}_i \boldsymbol{A}_i \widehat{\boldsymbol{\Psi}} \boldsymbol{A}_i^\top$. Additionally we can assume that $\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1}$ is finite, meaning that $\boldsymbol{b}_i$ lies in the span of the singular vectors of $\widehat{\boldsymbol{\Sigma}}_{b_i}$. (If this were not the case, the cost would be driven to infinity and we could not be at a minimizing solution anyway.) If $\{\widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\nu}}\}$ is a local minimum, then $\{\lambda_1 = 1, \lambda_2 = 0\}$ must be a local minimum of the revised cost function

$$\begin{aligned} \mathcal{L}(\lambda_1, \lambda_2) \;=\; & \sum_{i=1}^{m} \boldsymbol{b}_i^\top \left(\lambda_1 \widehat{\boldsymbol{\Sigma}}_{b_i} + \lambda_2 \boldsymbol{b}_i \boldsymbol{b}_i^\top \right)^{-1} \boldsymbol{b}_i \\ & + \log \left| \lambda_1 \widehat{\boldsymbol{\Sigma}}_{b_i} + \lambda_2 \boldsymbol{b}_i \boldsymbol{b}_i^\top \right|. \end{aligned} \tag{26}$$

This is because $\boldsymbol{b}_i \boldsymbol{b}_i^\top$ represents a valid set of basis vectors for updating the covariance per the construction above involving $\boldsymbol{\Psi}_i'$. First consider optimization over λ_1. If $\lambda_1 = 1$ is a local minimum, then by taking gradients and equating to zero, we require that

$$\sum_{i=1}^{m} \boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i = \sum_{i=1}^{m} \text{rank}\left[\widehat{\boldsymbol{\Sigma}}_{b_i} \right]. \tag{27}$$

Likewise, taking the gradient with respect to λ_2 we obtain

$$\left. \frac{\partial \mathcal{L}(\lambda_1, \lambda_2)}{\partial \lambda_2} \right|_{\lambda_1 = 1, \lambda_2 = 0} = \sum_{i=1}^{m} \boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i - \sum_{i=1}^{m} \left(\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i \right)^2. \tag{28}$$

The nullspace condition (a very mild assumption) ensures that $\sum_{i=1}^{m} \text{rank}[\widehat{\boldsymbol{\Sigma}}_{b_i}] = k$ for some $k > m$ when $\text{rank}[\boldsymbol{\Psi}] > 1$. To see this, observe that to achieve $\sum_{i=1}^{m} \text{rank}[\widehat{\boldsymbol{\Sigma}}_{b_i}] = m$ when $\text{rank}[\boldsymbol{\Psi}] > 1$ requires that $\boldsymbol{\Psi} = \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{W}\boldsymbol{W}^\top$ where $\boldsymbol{u}$ is a

vector and $\boldsymbol{W}$ is a matrix (or vector) with columns in $\mathrm{null}[\boldsymbol{A}_i]$, $\forall i$. If any such $\boldsymbol{W}$ is not in this nullspace for some i, then given that $p_i > 1$, the associated $\boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top$ will have rank greater than one, and the overall rank sum will exceed m.

Consequently, (28) will always be negative. This is because if $\sum_{i=1}^m z_i = k$ for any set of non-negative variables $\{z_i\}$, the minimal value of $\sum_{i=1}^m z_i^2$ occurs when $z_i = k/m$, $\forall i$. In our case, this implies that

$$\sum_{i=1}^{m} \left(\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i \right)^2 \geq \sum_{i=1}^{m} (k/m)^2 > k > m. \tag{29}$$

Therefore we can add a small contribution of $\boldsymbol{b}_i \boldsymbol{b}_i^\top$ to each $\widehat{\boldsymbol{\Sigma}}_{b_i}$ and reduce the underlying cost function. Hence we cannot have a local minimum, except when $\boldsymbol{\Psi}$ is equal to some $\boldsymbol{\Psi}^*$ with $\mathrm{rank}[\boldsymbol{\Psi}^*] = 1$. Moreover, we may directly conclude that $\boldsymbol{x}^* = \overline{\boldsymbol{\Psi}}^* \boldsymbol{A}^\top (\boldsymbol{A} \overline{\boldsymbol{\Psi}}^* \boldsymbol{A}^\top)^\dagger \boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*] = 1$ with $\boldsymbol{x}^* = \mathrm{vec}[\boldsymbol{X}^*]$.

Regarding the last part of the theorem, we consider only f that are concave non-decreasing functions (this is the only reasonable choice for shrinking singular values to zero, and the more general case naturally follows anyway with additional effort, but minimal enlightenment). Without loss of generality we may also assume that $f(0) = 0$ and $f(1) = 1$; we can always apply an inconsequential translation and scaling such that these conditions hold.[9] Simple counter examples then demonstrate that $f(\epsilon)$ must be greater than some constant C independent of ϵ for all ϵ sufficiently small. To see this, note that we can always rescale elements of $\boldsymbol{A}$ such that a solution with rank greater than one is preferred unless this condition holds. However, such an f, which effectively must display infinite gradient at $f(0)$ to guarantee a global solution is always rank one, will then always display local minima for certain $\boldsymbol{A}$. This can easily be revealed through simple counter-examples. ■

C. Symmetrization Update Rules

These iterative update rules follow from alternative upper bounds tailored to the symmetric version of BARM. When both $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_c$ are fixed, $\boldsymbol{x}$ is updated via the posterior mean calculation

$$\begin{aligned} \widehat{\boldsymbol{x}} = \ \mathrm{vec}\left[\widehat{\boldsymbol{X}}\right] &= \frac{1}{2}\left(\overline{\boldsymbol{\Psi}}_r + \overline{\boldsymbol{\Psi}}_c\right) \boldsymbol{A}^\top \\ &\times \left[\lambda \boldsymbol{I} + \boldsymbol{A} \frac{1}{2}\left(\overline{\boldsymbol{\Psi}}_r + \overline{\boldsymbol{\Psi}}_c\right) \boldsymbol{A}^\top\right]^{-1} \boldsymbol{b}. \end{aligned} \tag{30}$$

where $\overline{\boldsymbol{\Psi}}_r = \boldsymbol{\Psi}_r \otimes \boldsymbol{I}$ and $\overline{\boldsymbol{\Psi}}_c = \boldsymbol{I} \otimes \boldsymbol{\Psi}_c$. Likewise we update $\nabla_{\Psi_r^{-1}}$ and $\nabla_{\Psi_c^{-1}}$ using

$$\nabla_{\Psi_r^{-1}} = \sum_{i=1}^{m} \boldsymbol{\Psi}_r - \boldsymbol{\Psi}_r \boldsymbol{A}_{\mathrm{ri}}^\top \left(\boldsymbol{A} \overline{\boldsymbol{\Psi}}_r \boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_{\mathrm{ri}} \boldsymbol{\Psi}_r, \tag{31}$$

$$\nabla_{\Psi_c^{-1}} = \sum_{i=1}^{n} \boldsymbol{\Psi}_c - \boldsymbol{\Psi}_c \boldsymbol{A}_{\mathrm{ci}}^\top \left(\boldsymbol{A} \overline{\boldsymbol{\Psi}}_c \boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_{\mathrm{ci}} \boldsymbol{\Psi}_c, \tag{32}$$

[9]The log function is a limiting case, but what follows holds nonetheless.

where $\boldsymbol{A}_{\mathrm{ri}} \in \mathbb{R}^{p \times m}$ is defined such that $\boldsymbol{A} = [\boldsymbol{A}_{r1}^\top, \ldots, \boldsymbol{A}_{\mathrm{rm}}^\top]^\top$ and $\boldsymbol{A}_{\mathrm{ci}} \in \mathbb{R}^{p \times m}$ is defined such that $\boldsymbol{A} = [\boldsymbol{A}_{c1}, \ldots, \boldsymbol{A}_{\mathrm{cn}}]$. Finally given these values, with $\boldsymbol{X}$, $\nabla_{\Psi_r^{-1}}$ and $\nabla_{\Psi_c^{-1}}$ fixed, we can compute the optimal $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_c$ in closed form by optimizing the relevant $\boldsymbol{\Psi}_r$- and $\boldsymbol{\Psi}_c$-dependent terms via

$$\boldsymbol{\Psi}_r^{\mathrm{opt}} = \frac{1}{n}\left[\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}} + \nabla_{\Psi_r^{-1}}\right], \tag{33}$$

$$\boldsymbol{\Psi}_c^{\mathrm{opt}} = \frac{1}{m}\left[\widehat{\boldsymbol{X}} \widehat{\boldsymbol{X}}^\top + \nabla_{\Psi_c^{-1}}\right]. \tag{34}$$

In practice the simple initialization $\boldsymbol{\Psi}_r = \boldsymbol{I}$ and $\boldsymbol{\Psi}_c = \boldsymbol{I}$ is sufficient for obtaining good performance.

REFERENCES

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[2] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 35, no. 9, pp. 2117–2130, 2013.

[3] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 35, no. 1, pp. 171–184, 2013.

[4] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *Int. J. Comput. Vis. (IJCV)*, vol. 99, no. 1, pp. 1–24, 2012.

[5] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[6] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res. (JMLR)*, vol. 13, no. 1, pp. 3441–3473, 2012.

[7] Z. Li, J. Liu, Y. Jiang, J. Tang, and H. Lu, "Low rank metric learning for social image retrieval," in *Pro. 20th ACM Int. Conf. Multimedia*, 2012, pp. 853–856.

[8] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[9] B. Xin and D. Wipf, "Pushing the limits of affine rank minimization by adapting probabilistic pca," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 419–427.

[10] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[11] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, 2012.

[12] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, 2011.

[13] D. Wipf, "Non-convex rank minimization via an empirical Bayesian approach," in *Proc. 28th Conf. Uncertainty Artif. Intell. (UAI)*, 2012.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[15] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res. (JMLR)*, vol. 1, pp. 211–244, 2001.

[16] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

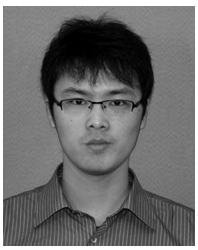
[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[18] E. J. Candes and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, 2014.

[19] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, 2015.

[20] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, NJ, USA: Prentice-Hall, 1969.

[21] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," *SIAM J. Scientif. Comput.*, vol. 35, no. 5, pp. S104–S125, 2013.
[22] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 937–945.
[23] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Found. Comput. Math.*, vol. 11, no. 2, pp. 183–210, 2011.
[24] R. H. Keshavan and S. Oh, *"A gradient descent algorithm on the Grassman manifold for matrix completion,"* 2009 DOI: arXiv Preprint arXiv:0910.5260.
[25] F. Léger, G. Yu, and G. Sapiro, *"Efficient matrix completion with Gaussian models,"* 2010 DOI: arXiv Preprint arXiv:1010.4050.
[26] B. Marlin, *Collaborative filtering: A machine learning perspective*, Ph.D. dissertation, Univ. of Toronto, Toronto, Canada ON, 2004.
[27] B. M. Marlin, "Modeling user rating profiles for collaborative filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003.
[28] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd ACM Int. Conf. Mach. Learn.*, 2005, pp. 713–719.
[29] S.-T. Park and D. M. Pennock, "Applying collaborative filtering techniques to movie search for better ranking and browsing," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2007, pp. 550–559.
[30] D. DeCoste, "Collaborative prediction using ensembles of maximum margin matrix factorizations," in *Proc. 23rd ACM Int. Conf. Mach. Learn.*, 2006, pp. 249–256.
[31] N. D. Lawrence and R. Urtasun, "Non-linear matrix factorization with gaussian processes," in *Proc. 26th Annu. ACM Int. Conf. Mach. Learn.*, 2009, pp. 601–608.
[32] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin, "Nonparametric Bayesian matrix completion," *Proc. IEEE SAM*, 2010.



Bo Xin (M'XX) received the B.S. degree in electronic engineering from Dalian University of Technology, China, in 2011. He is currently working toward the Ph.D. degree in computer science at Peking University, China. His research interests include optimization, machine learning and computer vision.

David Wipf (M'XX) received the B.S. degree with highest honors from the University of Virginia, and the Ph.D. degree from UC San Diego, where he was an NSF IGERT Fellow. Later he was an NIH Postdoctoral Fellow at UC San Francisco. Since 2011 he has been with Microsoft Research in Beijing. His research interests include Bayesian learning techniques applied to signal/image processing and computer vision. He is the recipient of several awards including the 2012 Signal Processing Society Best Paper Award, the Biomag 2008 Young Investigator Award, and the 2006 NIPS Outstanding Paper Award. Q2

Yizhou Wang (M'XX) received his Ph.D. in computer science from University of California at Los Angeles (UCLA) in 2005. He was a Research Staff of the Palo Alto Research Center (Xerox-PARC) from 2005 to 2008. He is currently a Professor of the Computer Science Department at Peking University (PKU), China. His research interests include computer vision, statistical modeling and learning.

Wen Gao (F'XX) received M.S. degree in computer science from Harbin Institute of Technology in 1985, and Ph.D. degree in electronics engineering from the University of Tokyo in 1991. He was a Professor in computer science at Harbin Institute of Technology from 1991 to 1995 and a Professor in computer science at Institute of Computing Technology of Chinese Academy of Sciences from 1996 to 2005. He is currently a Professor at the School of Electronics Engineering and Computer Science, Peking University, China. He has been leading research efforts to develop systems and technologies for video coding, face recognition, sign language recognition and synthesis, and multimedia retrieval. He earned many awards, which include five national awards for his research achievements and activities. He did many services to academic society, such as general co-chair of IEEE ICME07, and the head of Chinese delegation to the Moving Picture Expert Group (MPEG) of International Standard Organization (ISO). Since 1997, he is also the chairman of the working group responsible for setting a national Audio Video coding Standard (AVS) for China. He published four books and over 500 technical articles in refereed journals and proceedings in the areas of signal processing, image and video communication, computer vision, multimodal interface, pattern recognition, and bioinformatics. Q3

QUERIES

Q1. Author: For all conference paper references, provide page numbers if printed in proceeding or location of where conference was presented if not printed.

Q2. Author: Please provide a foward facing headshot for Dr. Wipf. Otherwise the bio will have to run without photo.

Q3. Author: Please provide initial year(s) of IEEE membership grade(s) for all authors.

Exploring Algorithmic Limits of Matrix Rank Minimization Under Affine Constraints

Bo Xin, *Member, IEEE*, David Wipf, *Member, IEEE*, Yizhou Wang, *Member, IEEE*, and Wen Gao, *Fellow, IEEE*

***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. Nonconvex 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 nonconvex 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.**

***Index Terms*—Rank minimization, affine constraints, matrix completion, matrix recovery, empirical Bayes.**

I. 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 [1]–[7]. Mathematically this involves solving

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

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

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

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 [1], [2]. 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

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

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 [1], [3], [6]. However, these guarantees typically only apply to a highly restrictive set of rank minimization problems, and in a practical setting nonconvex algorithms can succeed in a much broader range of conditions [2], [5], [6].

In Section II 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 III by the derivation of an alternative approach using Bayesian modeling techniques adapted from probabilistic PCA [8]. Section IV will then describe connections with nuclear norm minimization, convergence issues, and properties of global and local solutions. The latter includes special cases whereby any stationary point of the intrinsic cost function is guaranteed to have optimal rank, illustrating an underlying smoothing mechanism which leads to success over competing methods. We next discuss algorithmic enhancements in Section V that further improve recovery performance in practice. Section VI contains a wide variety of numerical comparisons that highlight the efficacy of this algorithm, while Section Section VII presents a computer vision application involving image rectification and a standard collaborative filtering benchmark. Technical proofs and algorithm update rule details are contained in the Appendix. Portions of this work have previously appeared in conference proceedings [9].

Manuscript received October 22, 2014; revised March 9, 2015 and November 23, 2015; accepted February 26, 2016. Date of publication April 7, 2016; date of current version. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tareq Al-Naffouri. The authors would like to thank the support from the following grants: 973-2015CB351800, NSFC-61231010, NSFC-61527804, NSFC-61210005 and the Microsoft Research Asia Collaborative Research funding.

B. Xin, Y. Wang, and W. Gao are with the Department of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jimxinbo@gmail.com; yizhou.wang@pku.edu.cn; wgao@pku.edu.cn).

D. Wipf is with the Visual Computing group, Microsoft Research, Beijing 100080, China (e-mail: davidwip@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2551697

Before proceeding, we highlight several main contributions as follows:

1) 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. Model justification is ultimately based on such meticulous technical considerations rather than merely the presumed qualitative legitimacy of any underlying prior distributions.
2) Non-convex algorithms have demonstrated some improvement in estimation accuracy over the celebrated convex nuclear norm; however, this typically requires the inclusion of one or more additional tuning parameters to incrementally inject additional objective function curvature and avoid bad local solutions. In contrast, for solving (1) our non-convex Bayesian-inspired algorithm requires no such parameters at all, and noisy relaxations necessitate only a single, standard trade-off parameter balancing data-fit and minimal rank.[1]
3) Over a wide battery of controlled experiments with ground-truth data, our approach outperforms all existing algorithms that we are aware of, Bayesian, non-convex, 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.

II. 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.

A. General Non-Convex Methods

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, [6] 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 $f(z) = \log(z^2 + \gamma)$, which maintains an intimate connection with rank given that

$$\log z = \lim_{q \to 0} q^{-1} \left(z^q - 1\right) \quad \text{and} \quad \lim_{q \to 0} z^q = I\left[z \neq 0\right], \quad (4)$$

where I is a standard indicator function. Consequently, when γ is small, $\sum_i \log(\sigma_i[\boldsymbol{X}]^2 + \gamma)$ behaves much like a scaled and translated version of the rank, albeit with nonzero gradients away from zero.

The IRLS0 algorithm from [6] 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, [5] 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 [2]. 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 [5] show that default settings perform relatively poorly.

Finally, a somewhat different class of non-convex algorithms can be derived using a straightforward application of alternating minimization [10]. 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

$$\min_{\boldsymbol{U},\boldsymbol{V}} \| b - \mathcal{A}\left(\boldsymbol{U}\boldsymbol{V}^T\right) \|_{\mathcal{F}} \quad (5)$$

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

[1] While not our emphasis here, similar to other Bayesian frameworks, even this trade-off parameter can ultimately be learned from the data if a true, parameter-free implementation is desired across noise levels.

matrices such that prior knowledge of all individual ranks is not feasible.

B. Bayesian Methods

From a probabilistic perspective, previous work has applied Bayesian formalisms to rank minimization problems, although not specifically within an affine constraint set. For example, [11]–[13] derive robust PCA algorithms built upon the linear summation of a rank penalty and an element-wise sparsity penalty. In particular, [12] applies an MCMC sampling approach for posterior inference, but the resulting iterations are not scalable, subjectable to detailed analysis, nor readily adaptable to affine constraints. In contrast, [11] applies a similar probabilistic model but performs inference using a variational mean-field approximation. While the special case of matrix completion is considered, from an empirical standpoint its estimation accuracy is not competitive with the state-of-the-art non-convex algorithms mentioned above. Finally, without the element-wise sparsity component intrinsic to robust PCA (which is not our focus here), [13] simply collapses to a regular PCA model with a closed-form solution, so the challenges faced in solving (1) do not apply. Consequently, general affine constraints really are a key differentiating factor.

From a motivational angle, the basic probabilistic model with which we begin our development can be interpreted as a carefully re-parameterized generalization of the probabilistic PCA model from [8]. This will ultimately lead to a non-convex algorithm devoid of the heuristic tuning strategies mentioned above, but nonetheless still uniformly superior in terms of estimation accuracy. We emphasize that, although we employ a Bayesian entry point for our algorithmic strategy, final justification of the underlying model will be entirely based on properties of the underlying cost function that emerges, rather than any putative belief in the actual validity of the assumed prior distributions or likelihood function. This is quite unlike the vast majority of existing Bayesian approaches.

C. Analytical Considerations

Turning to analytical issues, a number of celebrated theoretical results dictate conditions whereby substitution of the rank function with the convex nuclear norm in (1) is nonetheless guaranteed to still produce the minimal rank solution. For example, if $\mathcal{A}$ is a Gaussian iid measurement ensemble and $\boldsymbol{X}_0 \in \mathbb{R}^{n\times n}$ represents the optimal solution to (1) with $\mathrm{rank}[\boldsymbol{X}_0] = r$, then with high probability as the problem dimensions grow large, the minimum nuclear norm feasible solution will equal $\boldsymbol{X}_0$ if the number of measurements p satisfies $p \geq 3r(2n-r)$ [14].

The limitation of this type of result is two-fold. First, in the above situation the true minimum rank solution only actually requires $p \geq r(2n-r)$ measurements to be recoverable via brute force solution of (1), and the remaining difference of a factor of three can certainly be considerable in many practical situations (e.g., requiring 300 measurements is far more laborious than only needing 100 measurements). Secondly though, and far more importantly, all existing provable recovery guarantees place extremely strong restrictions on the structure of $\mathcal{A}$, e.g., strong restrictions on the singular value decay of matrices in the nullspace of $\mathcal{A}$. Such conditions are unlikely to ever hold in realistic application settings, including the image rectification example we describe in Section VII.A (in fact, these conditions are usually incapable of even being checked). In contrast, the algorithm we propose is empirically observed to only require the theoretically minimal number of measurements even when such nullspace conditions are violated in many cases. While a general theoretical guarantee of this sort is obviously not possible, we do nonetheless provide several supporting theoretical results indicative of why such performance is at least empirically obtainable.

III. Alternative Algorithm Derivation

In this section we first detail our basic distributional assumptions followed by development of the associated update rules for inference.

A. Basic Model

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

$$p\left(\boldsymbol{b}|\boldsymbol{X};\mathcal{A},\lambda\right) \quad \propto \quad \exp\left[-\frac{1}{2\lambda}\| \mathcal{A}\left(\boldsymbol{X}\right) - \boldsymbol{b} \|_2^2\right], \qquad (6)$$

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

$$p\left(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}\right) = \prod_i \mathcal{N}\left(\boldsymbol{x}_{:i};\boldsymbol{0},\nu_i\boldsymbol{\Psi}\right) \propto \exp\left[\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x}\right], \quad (7)$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{n\times n}$ is a positive semi-definite symmetric matrix,[2] $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_m]^\top$ is a non-negative vector, $\boldsymbol{x} = \mathrm{vec}[\boldsymbol{X}]$ (column-wise vectorization), and $\overline{\boldsymbol{\Psi}} = \mathrm{diag}[\boldsymbol{\nu}] \otimes \boldsymbol{\Psi}$, 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 (7) 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 $\widehat{\boldsymbol{X}}$ such that

$$\widehat{\boldsymbol{x}} = \mathrm{vec}\left[\widehat{\boldsymbol{X}}\right] = \overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\left(\lambda\boldsymbol{I} + \boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right)^{-1}\boldsymbol{b}. \qquad (8)$$

[2]Technically $\boldsymbol{\Psi}$ must be positive definite for the inverse in (7) to be defined. However, we can accommodate the semi-definite case using the following convention. Without loss of generality assume that $\overline{\boldsymbol{\Psi}} = \boldsymbol{R}\boldsymbol{R}^\top$ for some matrix $\boldsymbol{R}$. We then qualify that $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) = 0$ if $\boldsymbol{x} \notin \mathrm{span}[\boldsymbol{R}]$, and $p(\boldsymbol{X};\boldsymbol{\Psi},\boldsymbol{\nu}) \propto \exp[-\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}]$ otherwise. Equivalently, throughout the paper for convenience (and with slight abuse of notation) we define $\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \infty$ when $\boldsymbol{x} \notin \mathrm{span}[\boldsymbol{R}]$, and $\boldsymbol{x}^\top\overline{\boldsymbol{\Psi}}^{-1}\boldsymbol{x} = \boldsymbol{x}^\top(\boldsymbol{R}^\top)^\dagger\boldsymbol{R}^\dagger\boldsymbol{x}$ otherwise. This will come in handy, for example, when interpreting the bound in (12) below. Note also that the final cost function (10) we will ultimately be minimizing requires no such inverse anyway.

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 $\widehat{\boldsymbol{X}}$ 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 $\boldsymbol{\nu}$ [15], [13], [16]. This involves solving

$$\max_{\boldsymbol{\Psi} \in H^+, \boldsymbol{\nu} \geq 0} \int p\left(\boldsymbol{b} | \boldsymbol{X}; \mathcal{A}, \lambda\right) p\left(\boldsymbol{X}; \boldsymbol{\Psi}, \boldsymbol{\nu}\right) d\boldsymbol{X}, \tag{9}$$

where H^+ denotes the set of positive semi-definite and symmetric $n \times n$ matrices. After a $-2\log$ transformation and application of a standard convolution-of-Gaussians integration, solving (9) is equivalent to minimizing the cost function

$$\mathcal{L}\left(\boldsymbol{\Psi}, \boldsymbol{\nu}\right) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} + \log |\boldsymbol{\Sigma}_b|, \tag{10}$$

where

$$\boldsymbol{\Sigma}_b = \boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda \boldsymbol{I} \text{ and } \overline{\boldsymbol{\Psi}} = \operatorname{diag}\left[\boldsymbol{\nu}\right] \otimes \boldsymbol{\Psi}. \tag{11}$$

Here $\boldsymbol{\Sigma}_b$ is the covariance of $\boldsymbol{b}$ given $\boldsymbol{\Psi}$ and $\boldsymbol{\nu}$.

B. Update Rules

Minimizing (10) is a non-convex optimization problem, and we employ standard upper bounds for this purpose leading to an EM-like algorithm, somewhat related to [8]. 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 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 [16], for the first term in (10) we have

$$\boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \boldsymbol{x}^\top \overline{\boldsymbol{\Psi}}^{-1} \boldsymbol{x} \tag{12}$$

with equality whenever $\boldsymbol{x}$ satisfies (8). For the second term we use

$$\begin{aligned} \log |\boldsymbol{\Sigma}_b| &\equiv m \log |\boldsymbol{\Psi}| + \log |\lambda \boldsymbol{A}^\top \boldsymbol{A} + \overline{\boldsymbol{\Psi}}^{-1}| \\ &\leq m \log |\boldsymbol{\Psi}| + \operatorname{tr}\left[\boldsymbol{\Psi}^{-1} \nabla_{\Psi^{-1}}\right] + C, \end{aligned} \tag{13}$$

where because $\log |\lambda \boldsymbol{A}^\top \boldsymbol{A} + \overline{\boldsymbol{\Psi}}^{-1}|$ 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

$$\nabla_{\Psi^{-1}} = \sum_{i=1}^{m} \boldsymbol{\Psi} - \boldsymbol{\Psi} \boldsymbol{A}_i^\top \left(\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_i \boldsymbol{\Psi}, \tag{14}$$

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 (12) and (13) 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

$$\begin{aligned} \boldsymbol{\Psi}^{\text{opt}} &= \arg\min_{\boldsymbol{X}} \operatorname{tr}\left[\boldsymbol{\Psi}^{-1}\left(\boldsymbol{X}\boldsymbol{X}^\top + \nabla_{\Psi^{-1}}\right)\right] + m \log |\boldsymbol{\Psi}| \\ &= \frac{1}{m}\left[\widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top + \nabla_{\Psi^{-1}}\right]. \end{aligned} \tag{15}$$

By agnostically starting with $\boldsymbol{\Psi} = \boldsymbol{I}$ and then iteratively computing (8), (14), and (15), we can then obtain an estimate for $\boldsymbol{\Psi}$, and more importantly, a corresponding estimate for $\boldsymbol{X}$ given by (8) 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).

IV. Properties of BARM

Here we first describe a close but perhaps not intuitively-obvious relationship between the BARM objective function and canonical nuclear norm minimization. We then discuss desirable properties of global and local minima before concluding with a brief examination of convergence issues.

A. Connections with Nuclear Norm Minimization

On the surface, it may appear that minimizing (10) is completely unrelated to the convex problem

$$\min_{\boldsymbol{X}} \| \boldsymbol{X} \|_* \text{ s.t. } \boldsymbol{b} = \mathcal{A}\left(\boldsymbol{X}\right) \tag{16}$$

that is most commonly associated with practical rank minimization implementations. However, a close connection can be revealed by considering the modified objective function

$$\mathcal{L}'\left(\boldsymbol{\Psi}\right) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} + \operatorname{tr}\left[\overline{\boldsymbol{\Psi}}\right], \tag{17}$$

which represents nothing more than (10), with $\boldsymbol{\nu} = \boldsymbol{1}$ and with $\log |\boldsymbol{\Sigma}_b|$ being replaced by $\operatorname{tr}[\overline{\boldsymbol{\Psi}}]$. Now suppose we minimize (17) with respect to $\boldsymbol{\Psi} \in H^+$ obtaining some $\boldsymbol{\Psi}^*$. We then go on to compute an estimate of $\boldsymbol{X}$ using (8). Note that if we apply the bound from (12) to the first term in (17), then this estimate for $\boldsymbol{X}$ equivalently solves

$$\min_{\boldsymbol{\Psi} \in H^+, \boldsymbol{X}} \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + \boldsymbol{x}^\top \overline{\boldsymbol{\Psi}}^{-1} \boldsymbol{x} + \operatorname{tr}\left[\overline{\boldsymbol{\Psi}}\right], \tag{18}$$

with $\boldsymbol{x} = \operatorname{vec}[\boldsymbol{X}]$ as before. If we first optimize over $\boldsymbol{\Psi}$, it is easily demonstrated that the optimal value of $\boldsymbol{\Psi}$ equals $\left(\boldsymbol{X}\boldsymbol{X}^\top\right)^{1/2}$. Plugging this value into (18), simplifying, and then applying the definition of the nuclear norm, we arrive at

$$\min_{\boldsymbol{X}} \frac{1}{\lambda} \| \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \|_2^2 + 2 \| \boldsymbol{X} \|_*, \tag{19}$$

Furthermore, in the limit $\lambda \to 0$ (applied outside of the minimization), (19) becomes equivalent to (16). For more information regarding the duality relationship between variance/covariance space and coefficient space, at least in the related context of compressive sensing models, please refer to [16].

Consequently, we may conclude that the central distinction between the proposed BARM cost function and nuclear norm minimization is an intrinsic $\mathcal{A}$-dependent penalty function

$\log|\boldsymbol{\Sigma}_b|$ which is applied in covariance space. In Section IV.B we will examine desirable properties of this non-convex substitution, highlighting our desire to treat the underlying BARM probabilistic model as an independent cost function that may be subject to technical analysis independent of its Bayesian origins.

B. Global/Local Minima Analysis

As discussed in Section II one nice property of the $\sum_i \log(\sigma_i[\boldsymbol{X}])$ penalty employed (approximately) by IRLS0 [6] 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 $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ 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 $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ with $\gamma \to 0$. Before presenting these results, we define $\mathrm{spark}[\boldsymbol{A}]$ as the smallest number of linearly dependent columns in matrix $\boldsymbol{A}$ [17]. All proofs are deferred to the Appendix.

Lemma 1: Let $\boldsymbol{b}=\boldsymbol{A}\mathrm{vec}[\boldsymbol{X}]$, where $\boldsymbol{A}\in\mathbb{R}^{p\times \mathrm{nm}}$ 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 (10) in the limit $\lambda\to 0$ is such that $\boldsymbol{x}^*=\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top(\boldsymbol{A}\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top)^\dagger\boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*]=r$ with $\mathrm{vec}[\boldsymbol{X}^*]=\boldsymbol{x}^*$.

Lemma 2: Additionally, let $\widetilde{\boldsymbol{A}}=\boldsymbol{A}D$, 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 (10) in the limit $\lambda\to 0$, then $\{\boldsymbol{\Gamma}^{-1}\boldsymbol{\Psi}^*,\mathrm{diag}[\boldsymbol{\alpha}]^{-1}\boldsymbol{\nu}^*\}$ is a minimizer when $\widetilde{\boldsymbol{A}}$ 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 $p>\mathrm{rm}$. 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 II).

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 (9), 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 after some transformation $\boldsymbol{W}$, 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 $\cap_i\mathrm{null}[\boldsymbol{A}_i]=\varnothing$. Then if $\min_{\boldsymbol{X}}\mathrm{rank}[\boldsymbol{X}]=1$ in the feasible region, any minimizer $\{\boldsymbol{\Psi}^*,\boldsymbol{\nu}^*\}$ of (10) (global or local) in the limit $\lambda\to 0$ is such that $\boldsymbol{x}^*=\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top(\boldsymbol{A}\overline{\boldsymbol{\Psi}}^*\boldsymbol{A}^\top)^\dagger\boldsymbol{b}$ 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, at least in this rank-one case. (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.[3]

Although a global optimal guarantee for finding a rank-one matrix sounds somewhat limited, 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. Moreover, finding rank one matrices subject to affine constraints represents a crucial component of applications such as phase retrieval [18], [19].

Additionally, if a unique rank-one solution exists to (1), then the unique minimizing solution to (10) will produce this $\boldsymbol{X}$ via (8). 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.[4] Moreover, as evident from the experiments, the proposed algorithm always successfully finds the global optimal in many situations where the underlying matrix has a rank much higher than one. Therefore, although we can only provide theoretical guarantee for the rank-one case, the underlying intuition that local minima are smoothed away arguably carries over to situations where the rank is greater than one.

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

[4]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.

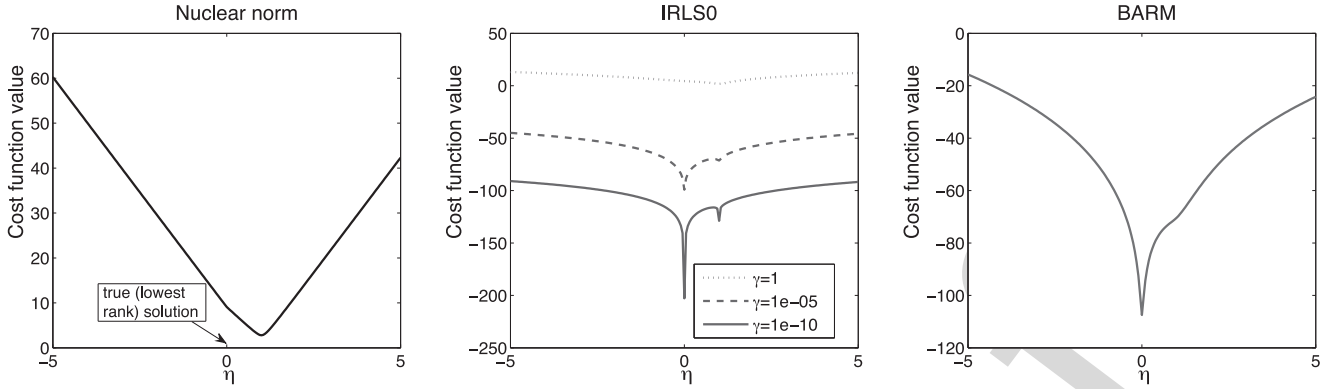


Fig. 1. Plots of different surrogates for matrix rank in a 1D feasible subspace. Here the convex nuclear norm does not retain the correct global minimum. In contrast, although the non-convex $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ penalty exhibits the correct minimum when γ is sufficiently small, it also contains spurious minima. Only BARM smoothes away local minimum while simultaneously retaining the correct global optima.

C. Visualization of BARM Local Minima Smoothing

To further explore the smoothing effect and complement Theorem 1, it helps to visualize rank penalty functions restricted to the feasible region. While the BARM algorithm involves minimizing (10), its implicit penalty function on $\boldsymbol{X}$ can nonetheless be numerically obtained across the feasible region in a given subspace of interest; for other penalties such as the nuclear norm this is of course trivial. Practically it is convenient to explore a 1D feasible subspace generated by $\boldsymbol{X}^*+\eta\boldsymbol{V}$, where $\boldsymbol{X}^*$ is the true minimum rank solution, $\boldsymbol{V}\in\text{null}[\boldsymbol{A}]$, and η is a scalar. We may then plot various penalty function values as η is varied, tracing the corresponding 1D feasible subspace. We choose $\boldsymbol{V}=\boldsymbol{X}^1-\boldsymbol{X}^*$, where $\boldsymbol{X}^1$ is a feasible solution with minimum nuclear norm; however, random selections from $\text{null}[\boldsymbol{A}]$ also show similar characteristics.

Fig. 1 provides a simple example of this process. $\boldsymbol{A}$ is generated randomly with all zeros and a single randomly placed '1' in each row leading to a canonical matrix completion problem. $\boldsymbol{X}^*\in\mathbb{R}^{5\times 5}$ is randomly generated as $\boldsymbol{X}^*=\boldsymbol{u}\boldsymbol{v}^\top$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are iid $\mathcal{N}(0,1)$ vectors, and so $\boldsymbol{X}^*$ is rank one. Finally, $p=10$ elements are observed, and therefore $\boldsymbol{A}$ has 10 rows and $5\times 5=25$ columns. η is varied from -5 to 5 and the values of the nuclear norm, $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$, and the implicit BARM cost function are displayed.

From the figure we observe that the minimum of the nuclear norm is not produced when the rank is smallest, which occurs when $\eta=0$; hence the convex cost function fails for this problem. Likewise, the $\sum_i \log(\sigma_i[\boldsymbol{X}]^2+\gamma)$ penalty used by IRLS0 displays an incorrect global minimum when the tuning parameter γ is large. In contrast, when γ is small, while the global minimum may now be correct, spurious local ditches have appeared in the cost function.[5] Therefore, any success of the IRLS0 algorithm depends heavily on a carefully balanced decaying sequence of γ values, with the hope that initial iterations can steer the trajectory towards a desirable basin of attraction where local minima are less problematic. One advantage of BARM then is that it is parameter free in this respect and yet still retains the correct global minimum, often without additional spurious local minima.

[5]Technically speaking, these are not provably local minima since we are only considering a 1D subspace of the feasible region. However, it nonetheless illustrates the strong potential for troublesome local minima, especially in high dimensional practical problems.

D. Convergence

Previous results of Section IV are limited to exploring aspects of the underlying BARM cost function. Regarding the BARM algorithm itself, by construction the updates generated by (8), (14), and (15) 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 [20]. However, provided we add a small regularization factor $\gamma\text{tr}[\boldsymbol{\Psi}^{-1}]$, 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 $\boldsymbol{\Psi}^*\in\mathcal{S}$, $\mathcal{L}(\boldsymbol{\Psi}^*)+\gamma\text{tr}[(\boldsymbol{\Psi}^*)^{-1}]$ 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 [6] as discussed in Section II. 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.

V. 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 $\boldsymbol{x} = \text{vec}[\boldsymbol{X}]$ 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

$$\overline{\boldsymbol{\Psi}} = 1/2\left(\boldsymbol{\Psi}_r \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{\Psi}_c\right). \tag{20}$$

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 listed in the Appendix and computational complexity evaluated in Section VI.E.

VI. Experimental Validation

This section compares BARM with existing state-of-the-art affine rank minimization algorithms. For BARM, in all noiseless cases we simply used $\lambda = 10^{-10}$ (effectively zero), and hence no tuning parameters are required. Likewise, nuclear norm minimization [1], [4] 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 [6]. 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 [5] where available. Note that both [5] and [6] 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 [10] or the normalized hard thresholding algorithm from [21]. 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. However, we do show limited empirical results with the variational sparse Bayesian algorithm (VSBL) from [11] 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.

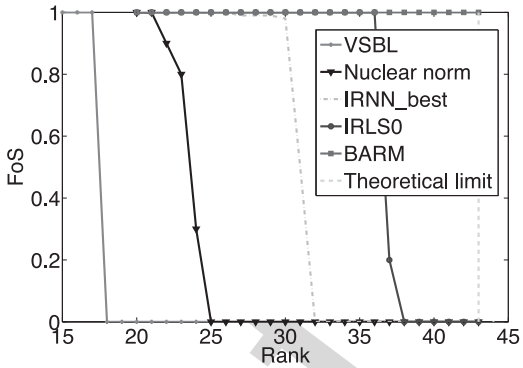


Fig. 2. Matrix completion comparisons (avg of 10 trials).

A. 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 [5], where a rank r matrix is generated as $\boldsymbol{X}_0 = \mathbf{TM}_L\mathbf{TM}_R$, with $\mathbf{TM}_L \in \mathbb{R}^{n\times r}$ and $\mathbf{TM}_R \in \mathbb{R}^{r\times m}$ $(n = m = 150)$ as iid $\mathcal{N}(0,1)$ random matrices. 50% of all entries are then hidden uniformly at random. The *relative error* (REL) given by $\| \boldsymbol{X}_0 - \widehat{\boldsymbol{X}}\|_{\mathcal{F}}/\| \boldsymbol{X}_0 \|_{\mathcal{F}}$ 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 Fig. 2 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 [5] 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 [6] 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 [6] tests a variety of

TABLE I
MATRIX COMPLETION RESULTS OF BARM WITH IRLS0 ON THE THREE HARDEST PROBLEMS FROM [6]. PUBLISHED RESULTS IN [6] 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

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 [6]. We also only include experiments where algorithms are blind to the true rank of $\boldsymbol{X}_0$.[6] Results are shown in Table I, where we have also displayed the published results of three additional algorithms that were compared with IRLS0 in [6], namely, IHT [22], FPCA [23] and Optspace [24]. From the table we observe that, in the most difficult problem considered in [6], IRLS0 achieved only a 0.5 FoS score (meaning failure 50% of the time) while BARM still achieves a perfect 1.0. Note that when FR is high, the problem of recovering the underlying matrix is essentially much harder. This happens in a manner that more local minima are induced (due to increased rank) and/or much larger search space are exposed (due to decreased number of observations/constraints). In these cases, the equivalency of the global optimal with convex relaxation usually does not hold, whereas for the existing non-convex surrogates, there is no reason to assume any local minima are not present. However, since BARM has an implicit mechanism of smoothing local minima (though maybe not all of them), it works more robustly in these situations.

B. General $\boldsymbol{A}$

Next we consider the more challenging problem involving arbitrary affine constraints. The desired low-rank $\mathbf{T}\mathbf{X}_0$ 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. Fig. 3 displays aggregate results when $\boldsymbol{X}_0$ is 50×50 and 100×100, including the underlying REL scores for additional comparison. 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.

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

C. Failure Case Analysis

Thus far we have not shown any cases where BARM actually fails. Of course solving (1) for general $\boldsymbol{A}$ 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 $p = 975$.

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. In Fig. 4 we have averaged the singular values of $\widehat{\boldsymbol{X}}$ in all the 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. 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 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 $\widehat{\boldsymbol{X}}$ is found such that $\sigma_r[\widehat{\boldsymbol{X}}]/\sigma_{r+1}[\widehat{\boldsymbol{X}}] > 10^3$, 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

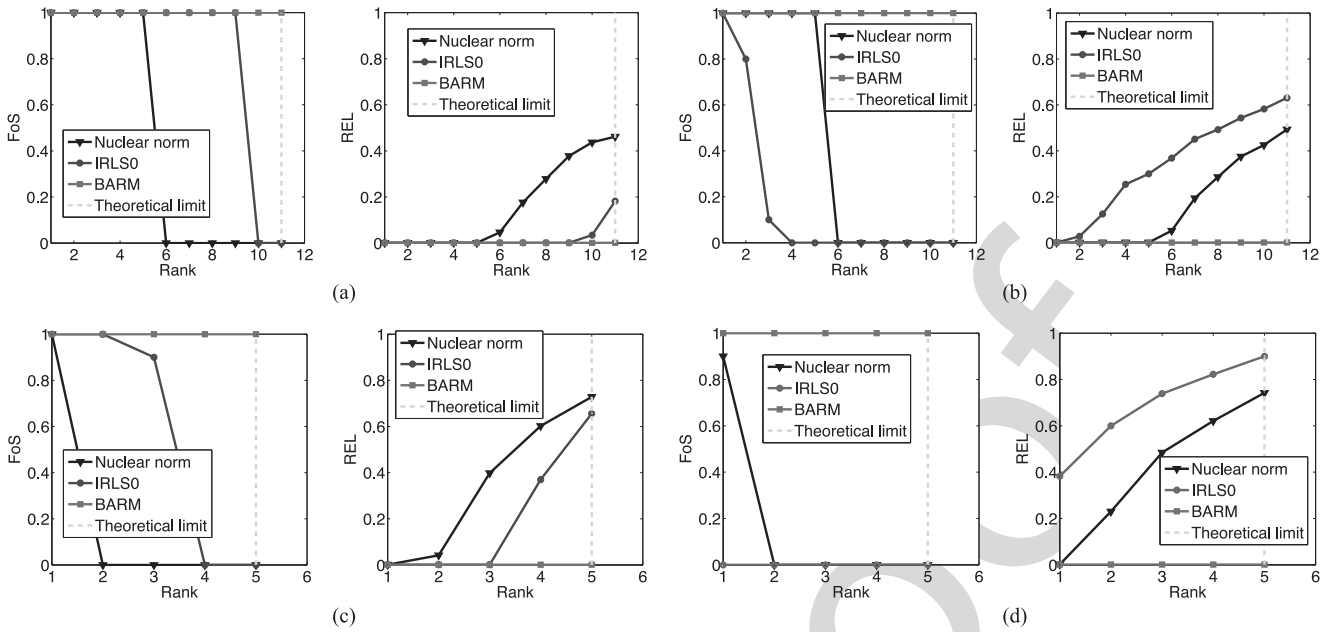


Fig. 3. Comparisons with general affine constraints (avg of 10 trials). (a) 50×50, $\boldsymbol{A}$ uncorrelated, (b) 50×50, $\boldsymbol{A}$ correlated, (c) 100×100, $\boldsymbol{A}$ uncorrelated, and (d)100×100, $\boldsymbol{A}$ correlated.

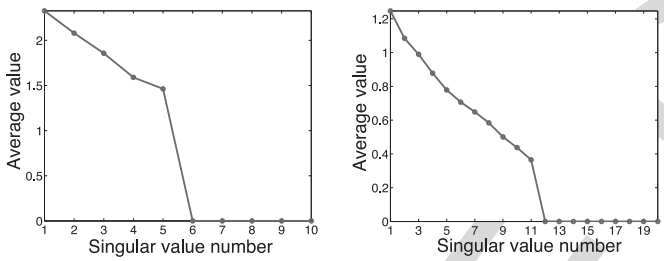


Fig. 4. Singular value averages of failure cases. In both cases solutions of minimal rank are obtained even though $\widehat{\boldsymbol{X}} \neq \boldsymbol{X}_0$. (a) 50×50 and (b) 100×100.

TABLE II
FURTHER MATRIX COMPLETION COMPARISONS OF BARM WITH IRLS0 BY REDUCING THE NUMBER OF MEASUREMENTS IN THE HARDEST PROBLEM FROM [6]. 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

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 I the most difficult case involves $\mathrm{FR} = 0.87$; 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 II. While IRLS0 (which is the top performing algorithm in [6] and in our experiments besides BARM) fails 100% of the time via both metrics, BARM can achieve an FoS of 0.7 even when $\mathrm{FR} = 0.99$ 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. Fig. 5 depicts the results using problem sizes consistent with Fig. 3; note that the FoRS metric has replaced the REL metric for comparison purposes.

Two things stand out from the 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 $\widehat{\boldsymbol{X}}$ 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

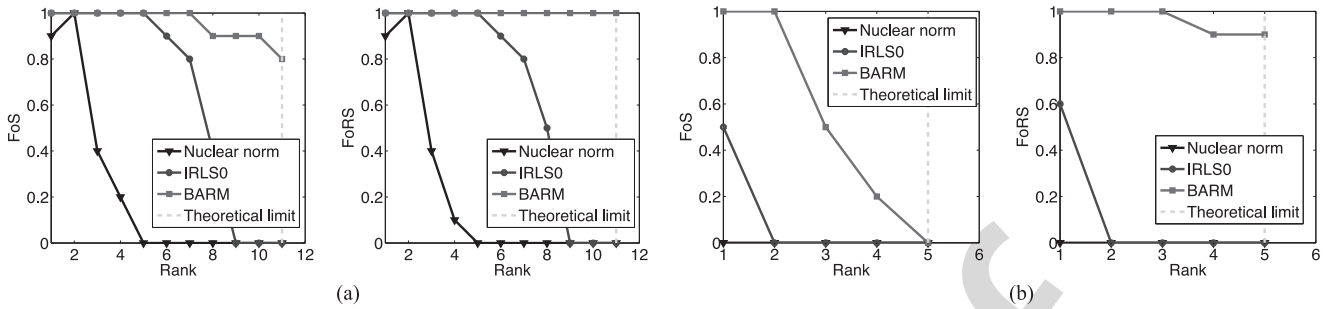


Fig. 5. Comparisons with structured affine constraints using both FoS and FoFS evaluation metrics (avg of 10 trials). (a) 50×50, $\boldsymbol{A}$ sub-sampled DCT, (b) 100×100, $\boldsymbol{A}$ sub-sampled DCT.

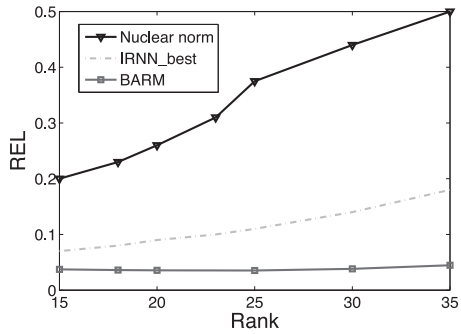


Fig. 6. Test with noisy data.

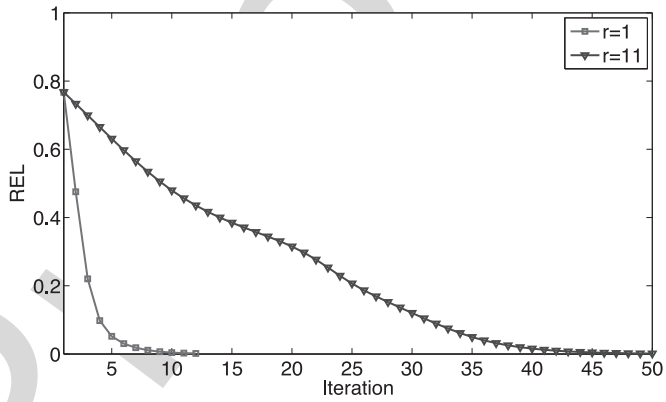


Fig. 7. Empirical convergence of BARM.

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.

D. Additional Noisy Tests

We also briefly present results that demonstrate the robustness of BARM to noise. For this purpose we reproduce the noisy experiment from [5] designed for validating IRNN algorithms. The simulated data are generated in the exact same way as was used to produce Fig. 2, only now instead of observing elements of $\boldsymbol{X}_0$ directly, we observe $\boldsymbol{X}_0 + 0.1 \times \boldsymbol{E}$, where elements of $\boldsymbol{E}$ are iid $\mathcal{N}(0,1)$. Although in [5] a heuristic strategy is introduced and tuned for adaptively setting all parameters (four in total), we simply applied BARM with $\lambda = 10^{-3}$ (so only a single parameter need be adjusted, and actually a wide range of λ values produces similar performance anyway). Results are shown in Fig. 6 where we compare BARM directly with the best result reported in [5] over the range $r = 15$ to $r = 35$. The nuclear norm solution is also included for reference. Overall, the BARM solution is stable and exhibits superior accuracy relative to the others.

E. Computational Complexity

Finally, regarding computational complexity, for general $\boldsymbol{A}$ the BARM updates can be implemented to scale linearly in the elements of $\boldsymbol{X}$ and quadratically in the number of observations p (the special case of matrix completion is decidedly much cheaper because of the special structure that can be exploited). In our experiments, for relatively easy problems on the order of 10 iterations are required, while for difficult recovery problems near the theoretical recovery boundary this may increase by a factor of 10 or so. This is somewhat expected though since as we near the theoretical limit, $\boldsymbol{A}$ becomes highly overcomplete, and candidate solutions become much more difficult to differentiate.

To show this effect empirically, we compare two separate trials from Fig. 3(a), the first when $r = 1$ (relatively easy), the second when $r = 11$ (relatively hard).[7] In Fig. 7 we plot the value of REL in both cases versus the iteration number of BARM.

VII. Application Examples

Many real-world problems from disparate fields can be formulated as the search for a low-rank matrix under affine constraints [1], [3], [4], [25]. Here we briefly consider two such examples: low-rank 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.

A. Low-Rank Image Rectification

In [4], the *transform invariant low-rank textures* (TILT) algorithm is derived for rectifying images containing low-rank

[7] Note that $r = 1$ is only relatively easy here because the number of observations is sufficient for the larger $r = 11$ case; if only the minimal number of measurements are available then even $r = 1$ can be challenging for many algorithms.

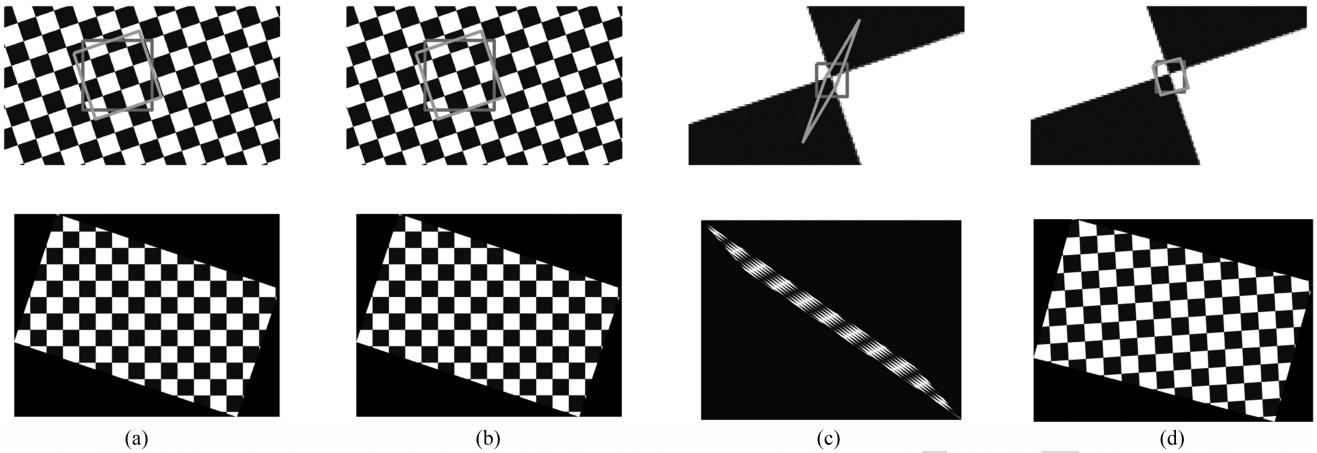


Fig. 8. Image rectification comparisons using a checkboard image. *Top*: Original image with observed region (red box) and estimated transformation (green box). *Bottom*: Rectified image estimates. (a) Nuclear norm (easy), (b) BARM (easy), (c) Nuclear norm (hard), (d) BARM (hard).

textures that have been transformed using an unknown operator $\boldsymbol{\tau}$ from some group (e.g., a homography). For a given observed image $\boldsymbol{Y}$, the basic idea is to construct a first-order Taylor series approximation around the current rectified image estimate $\widehat{\boldsymbol{X}}$ and solve

$$\min_{\boldsymbol{X},\boldsymbol{\delta}} \operatorname{rank}[\boldsymbol{X}] \text{ s.t. } \boldsymbol{X} = \boldsymbol{Y} + \sum_i \boldsymbol{J}_i\left(\widehat{\boldsymbol{X}}\right)\delta_i, \tag{21}$$

where $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$ is the Jacobian matrix with respect to $\boldsymbol{X}$ of the i-th parameter τ_i describing the transformation, with $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots]^\top$. Optimization over the vector of first-order differences $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots]^\top$ can be accomplished in closed form by projecting both sides of the constraint to the orthogonal complement of the span of all $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$. Let $P_{\boldsymbol{J}^c}$ represent this projection operator. The feasible region in (21) then becomes

$$P_{\boldsymbol{J}^c}(\boldsymbol{X}) = P_{\boldsymbol{J}^c}(\boldsymbol{Y}) + P_{\boldsymbol{J}^c}\left(\sum_i \boldsymbol{J}_i\left(\widehat{\boldsymbol{X}}\right)\delta_i\right) = P_{\boldsymbol{J}^c}(\boldsymbol{Y}) \tag{22}$$

The resulting problem then reduces exactly to (1) when we define $\mathcal{A} = P_{\boldsymbol{J}^c}$ and $\boldsymbol{b} = \operatorname{vec}[P_{\boldsymbol{J}^c}(\boldsymbol{Y})]$. Once $\boldsymbol{X}$ is computed in this way, we then update each $\boldsymbol{J}_i(\widehat{\boldsymbol{X}})$ and repeat until convergence.

While the original TILT algorithm substitutes the nuclear norm for $\operatorname{rank}[\boldsymbol{X}]$, we embedded the BARM algorithm into the posted TILT source code [4] for comparison purposes (note that we disabled an additional sparse error term for both algorithms to simplify comparisons, and it is not necessary anyway in many regimes). Figs. 8 and 9 display results on both two easy examples, where the number of observations p is large, and two more difficult problems where the number observations is small. While both algorithms succeed on the easy cases, when the observations are constrained by a small image window, only BARM is successful in accurately rectifying the images. This may be due, at least in part, to the fact that the implicit $\mathcal{A}$ operator contains significant structure that is not consistent with the required nullspace properties required for nuclear norm minimization success.

B. Collaborative Filtering of MovieLens Data

Collaborative filtering, a technique used by many recommender systems, is a popular representative application of low-rank matrix completion. Typically the rows (or columns) of $\boldsymbol{X}_0$ index users, the columns (or rows) denote items, and each entry $(\boldsymbol{X}_0)_{\text{ij}}$ is the rating/score of user i applied to item j. Given that we can observe some subset of elements of $\boldsymbol{X}_0$, the task of collaborative filtering is to predict all or some of the missing ratings. In general this would be impossible; however, if we have access to some prior knowledge, e.g., $\boldsymbol{X}_0$ is low-rank, then estimation may be feasible.

While our interest here is not in recommender systems or collaborative filtering per se, we nonetheless evaluate BARM using the 1M MovieLens dataset[8] as this appears to represent one of the most common evaluation benchmarks. We emphasize at the outset that the strict validity of any low-rank assumptions underlying this data is debatable, and it remains entirely unclear whether the true globally optimal or lowest rank solution consistent with the observations, even if computable, would necessarily lead to the best prediction of the unknown ratings. In fact, the reported performance of various existing rank-minimization algorithms tends to cluster around almost the same value, implying that collaborative filtering may not provide the most discriminative data type with which to compare. In most cases, it appears that tuning parameters and other heuristic modifications play a larger role than the underlying algorithmic distinctions fundamental to finding optimal low-rank estimates. Nonetheless, we apply BARM for completeness and convention, adopting an additional simple mean-offset estimation term from [25] that is particularly suitable for this problem.

In [6], IRLS0 is compared with only two other algorithms on MovieLens data, but the performance is no better. Therefore, we choose to compare directly with [25], which both derives an IRLS-like algorithm and shows comparisons with a much wider variety of alternative algorithms using a strict evaluation protocol that is standard in the literature. Specifically, the

[8]http://www.grouplens.org/

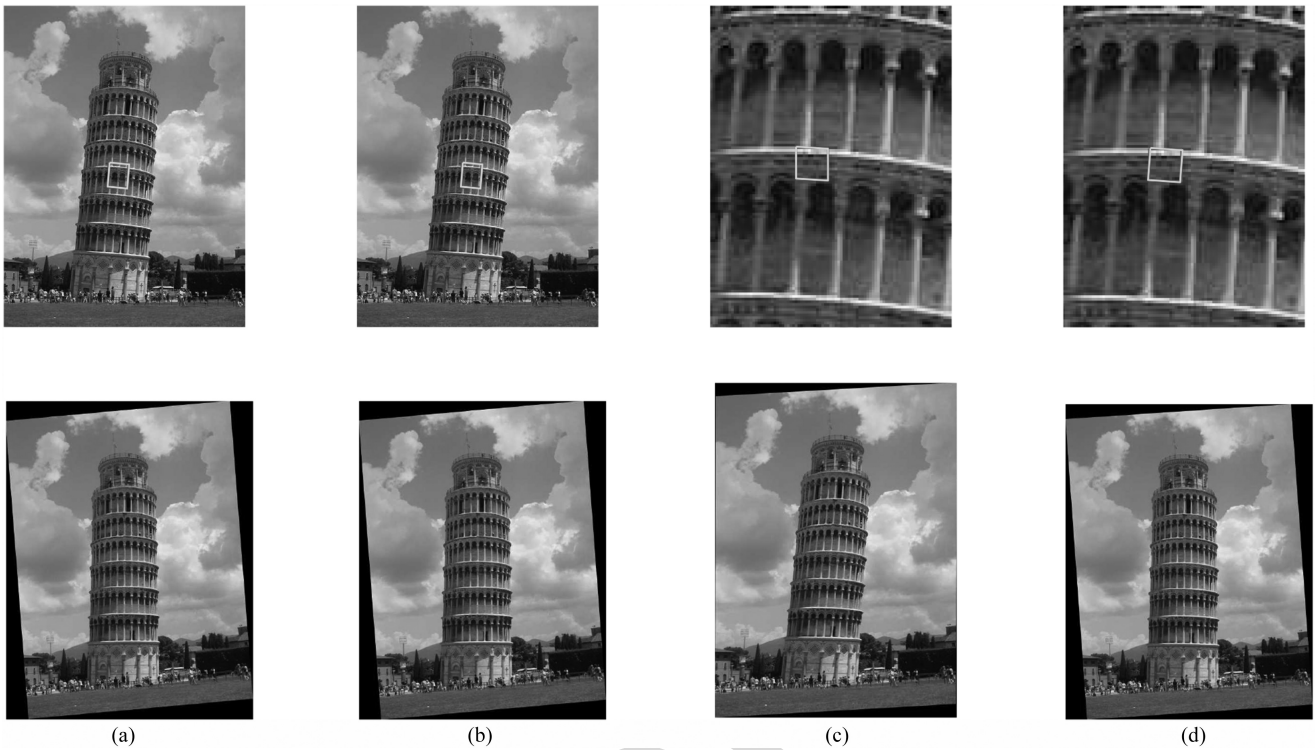


Fig. 9. Image rectification comparisons using a landmark photo. *Top*: Original image with observed region (red box) and estimated transformation (green box). *Bottom*: Rectified image estimates. (a) Nuclear norm (easy), (b) BARM (easy), (c) Nuclear norm (hard), (d) BARM (hard).

1M MovieLens dataset, which contains 1 million ratings in the range $\{1, \ldots, 5\}$ for 3900 movies from 6040 unique users, is assessed under two test-protocals: *weak generalization*, which measures the ability to predict other items rated by the same user, and *strong generalization*, which measures the ability to predict items by novel users. 5 000 users are randomly selected for the weak generalization, and likewise 1 000 users are extracted for the strong generalization. Each experiment is then run three times and the averaged results are reported. The performance metric is *normalized mean absolute error* (NMAE) given as

$$\text{NMAE} = \frac{\left(\sum_{i,j \in \text{supp}(\boldsymbol{X}_0)} \frac{|(\boldsymbol{X}_0)_{\text{ij}} - \widehat{\boldsymbol{X}}_{\text{ij}}|}{|\text{supp}(\boldsymbol{X}_0)|} \right)}{(rt_{\max} - rt_{\min})},$$

where $rt_{\max}$ and $rt_{\min}$ are the maximum and minimum ratings possible.

We followed the same setup and reported results using BARM in Table III along with results from [25] for comparison. This includes the additional algorithms URP [26], Attitude [27], MMMF [28], IPCF [29], E-MMMF [30], GPLVM [31], NBMC [32], and IRLS/GM [25], [6]. From this table we observe that for the easier weak generalization problem BARM is a close second best, while for the more challenging strong generalization BARM is actually the best. Of course it is also immediately apparent that all algorithms fall within a relatively narrow performance range of approximately five percentage points. Consequently, we cannot unequivocally conclude that the attributes of BARM which make it suitable for optimally minimizing rank necessarily translate into a truly significant practical advantage on this collaborative filtering task. But we would argue that the same holds for any matrix completion algorithm.

TABLE III
COLLABORATIVE FILTERING ON 1M MOVIELENS DATASET. RESULTS FROM [25] ARE IN ITALIC FOR COMPARISON PURPOSES

	Weak NMAE	Hard NMAE
URP	*0.4341*	*0.4444*
Attitude	*0.4320*	*0.4375*
MMMF	*0.4156*	*0.4203*
IPCF	*0.4096*	*0.4113*
E-MMMF	*0.4029*	*0.4071*
GPLVM	*0.4026*	*0.3994*
NBMC	**0.3916**	*0.3992*
IRLS/GM	*0.3959*	*0.3928*
BARM	0.3942	**0.3898**

VIII. CONCLUSION

This paper explores a conceptually-simple, parameter-free algorithm called BARM 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. Our strategy in this effort has been to adopt Bayesian machinery for inspiring a principled cost function; however, ultimate model justification is placed entirely in

theoretical evaluation of desirable global and local minima properties, and in the empirical recovery performance that inevitably results from these properties. Although in general non-convex algorithms are exponentially more challenging to analyze, in this regard we have at least attempted to contextualize BARM in the same manner as convex optimization-based approaches such as nuclear-norm minimization.

APPENDIX A

Here we provide brief proofs of Lemmas 1 and 2 as well as Theorem 1. We also address the augmented update rules that account for the revised, symmetrized cost function discussed in Section V.

A. Proof of Lemmas 1 and 2

Regarding Lemma 1, this result mirrors related ideas from [16] in the context of Bayesian compressive sensing. Hence, while a more rigorous presentation is possible, here we describe the basic aspects of the adaptation. At any candidate minimizer of (10) in the limit $\lambda \to 0$, define $\boldsymbol{W}$ such that $\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top = \boldsymbol{W}\boldsymbol{W}^\top$. To be a minimizer, global or local, it must be that $\boldsymbol{b} \in \text{span}[\boldsymbol{W}]$. If this were not the case, then $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ would diverge to infinity as $\lambda \to 0$ because $\boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b}$ progresses to infinity at a faster rate than $\log|\boldsymbol{\Sigma}_b|$ can compensate by approaching minus infinity. Intuitively, in much the same way $\text{argmin}_z \frac{1}{z} + \log z = 1$, meaning the optimal z must lie in the 'span' of 1 else the overall objective will be driven to infinity.

Consequently, the only way to minimize the cost in the limit as $\lambda \to 0$ is to consider low-rank solutions within the constraint set that $\boldsymbol{b} \in \text{span}[\boldsymbol{W}]$, and it is equivalent to requiring that $\boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq C$ for some constant C independent of λ (which ultimately corresponds with maintaining $\mathcal{A}(\boldsymbol{X}) = \boldsymbol{b}$ in the limit as well).

In this setting, while $0 \leq \boldsymbol{b}^T \boldsymbol{\Sigma}_b^{-1} \boldsymbol{b} \leq C$ is bounded, the second term in $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ can be unbounded from below when $\text{rank}[\boldsymbol{\Psi}]$ is sufficiently small. To see this note that

$$\log|\boldsymbol{\Sigma}_b| = \sum_{i=1}^{p} \log\left(\sigma_i\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] + \lambda\right), \tag{23}$$

where $\sigma_i[\cdot]$ denotes the i-th singular value of a matrix. While the maximum rank of $\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top$ is obviously p, if $r \triangleq \text{rank}[\boldsymbol{\Psi}] < p/m$ and $\text{spark}[\boldsymbol{A}] = p+1$ (maximal spark) as stipulated in the lemma statement, then $\text{rank}\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] = mr$ and (23) becomes

$$\log|\boldsymbol{\Sigma}_b| = \sum_{i=1}^{mr} \log\left(\sigma_i\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] + \lambda\right) + (p - mr)\log\lambda. \tag{24}$$

Note that the spark assumption accomplishes two objectives in this context. First, it guarantees that a high rank $\boldsymbol{\Psi}$ cannot masquerade as a low rank $\boldsymbol{\Psi}$ behind the nullspace of some collection of columns $\boldsymbol{A}_i$. Secondly, it ensures that after assuming $r < p/m$, then $\text{rank}\left[\boldsymbol{A}\overline{\boldsymbol{\Psi}}\boldsymbol{A}^\top\right] = mr$.

Consequently, in the limit where $\lambda \to 0$ (with the limit being taken outside of the minimization), (23) effectively scales as $(p - mr)\log\lambda$, and hence the overall cost is minimized when $\boldsymbol{\Psi}$ has minimal rank. This in turn ensures that the corresponding $\boldsymbol{X}$ will also have minimal rank, completing the proof sketch for Lemma 1.

Finally, Lemma 2 follows directly from the structure of the $\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu})$ cost function via simple reparameterizations. ■

B. Proof of Theorem 1

To begin we assume that $\boldsymbol{b}_i \neq 0$, $\forall i$, where $\boldsymbol{b}_i$ denotes the sub-vector of $\boldsymbol{b}$ such that $\boldsymbol{b}_i = \boldsymbol{A}_i \boldsymbol{x}_{:i}$. If this were not the case we can always collapse $\boldsymbol{X}$ by the corresponding column (which is indistinguishable from zero) and achieve an equivalent result. Given the assumptions of Theorem 1, the BARM cost function becomes

$$\mathcal{L}(\boldsymbol{\Psi}, \boldsymbol{\nu}) = \sum_{i=1}^{m} \boldsymbol{b}_i^\top \left(\nu_i \boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top\right)^{-1} \boldsymbol{b}_i + \log\left|\nu_i \boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top\right|. \tag{25}$$

If there exists a feasible rank one solution to $\boldsymbol{b} = \boldsymbol{A}\text{vec}[\boldsymbol{X}]$, then there also exists a set of $\boldsymbol{\Psi}_i' = \nu_i \boldsymbol{\Psi}$ such that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \boldsymbol{A}_i \boldsymbol{\Psi}_i' \boldsymbol{A}_i^\top$ for all i. To see this, note that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \boldsymbol{A}_i \boldsymbol{x}_{:i} \boldsymbol{x}_{:i}^\top \boldsymbol{A}_i^\top$. Because $\text{rank}[\boldsymbol{X}] = 1$, it also follows that $\boldsymbol{b}_i \boldsymbol{b}_i^\top = \alpha_i \boldsymbol{A}_i \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{A}_i^\top$, where $\alpha_i = \| \boldsymbol{x}_{:i} \boldsymbol{x}_{:i}^\top \| / \| \boldsymbol{X}\boldsymbol{X}^\top \|$. Therefore $\boldsymbol{\Psi}_i' = \nu_i \boldsymbol{X}\boldsymbol{X}^\top$ achieves the desired result with $\nu_i = \alpha_i$.

Now suppose we have converged to any solution $\{\widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\nu}}\}$ with $\text{rank}[\boldsymbol{\Psi}] > 1$ and associated $\widehat{\overline{\boldsymbol{\Psi}}} = \boldsymbol{I} \otimes \widehat{\boldsymbol{\Psi}}$. Note that since $\boldsymbol{b}_i \neq 0$, $\boldsymbol{\nu}_i > 0$ for all i, otherwise a local minimum is not possible (the cost function would be driven to positive infinity).

Define $\widehat{\boldsymbol{\Sigma}}_{b_i} = \hat{\nu}_i \boldsymbol{A}_i \widehat{\boldsymbol{\Psi}} \boldsymbol{A}_i^\top$. Additionally we can assume that $\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1}$ is finite, meaning that $\boldsymbol{b}_i$ lies in the span of the singular vectors of $\widehat{\boldsymbol{\Sigma}}_{b_i}$. (If this were not the case, the cost would be driven to infinity and we could not be at a minimizing solution anyway.) If $\{\widehat{\boldsymbol{\Psi}}, \widehat{\boldsymbol{\nu}}\}$ is a local minimum, then $\{\lambda_1 = 1, \lambda_2 = 0\}$ must be a local minimum of the revised cost function

$$\begin{aligned}\mathcal{L}(\lambda_1, \lambda_2) \;=\; & \sum_{i=1}^{m} \boldsymbol{b}_i^\top \left(\lambda_1 \widehat{\boldsymbol{\Sigma}}_{b_i} + \lambda_2 \boldsymbol{b}_i \boldsymbol{b}_i^\top\right)^{-1} \boldsymbol{b}_i \\ & + \log\left|\lambda_1 \widehat{\boldsymbol{\Sigma}}_{b_i} + \lambda_2 \boldsymbol{b}_i \boldsymbol{b}_i^\top\right|. \end{aligned} \tag{26}$$

This is because $\boldsymbol{b}_i \boldsymbol{b}_i^\top$ represents a valid set of basis vectors for updating the covariance per the construction above involving $\boldsymbol{\Psi}_i'$. First consider optimization over λ_1. If $\lambda_1 = 1$ is a local minimum, then by taking gradients and equating to zero, we require that

$$\sum_{i=1}^{m} \boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i = \sum_{i=1}^{m} \text{rank}\left[\widehat{\boldsymbol{\Sigma}}_{b_i}\right]. \tag{27}$$

Likewise, taking the gradient with respect to λ_2 we obtain

$$\left.\frac{\partial \mathcal{L}(\lambda_1, \lambda_2)}{\partial \lambda_2}\right|_{\lambda_1 = 1, \lambda_2 = 0} = \sum_{i=1}^{m} \boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i - \sum_{i=1}^{m} \left(\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i\right)^2. \tag{28}$$

The nullspace condition (a very mild assumption) ensures that $\sum_{i=1}^{m} \text{rank}[\widehat{\boldsymbol{\Sigma}}_{b_i}] = k$ for some $k > m$ when $\text{rank}[\boldsymbol{\Psi}] > 1$. To see this, observe that to achieve $\sum_{i=1}^{m} \text{rank}[\widehat{\boldsymbol{\Sigma}}_{b_i}] = m$ when $\text{rank}[\boldsymbol{\Psi}] > 1$ requires that $\boldsymbol{\Psi} = \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{W}\boldsymbol{W}^\top$ where $\boldsymbol{u}$ is a

vector and $\boldsymbol{W}$ is a matrix (or vector) with columns in $\mathrm{null}[\boldsymbol{A}_i]$, $\forall i$. If any such $\boldsymbol{W}$ is not in this nullspace for some i, then given that $p_i > 1$, the associated $\boldsymbol{A}_i \boldsymbol{\Psi} \boldsymbol{A}_i^\top$ will have rank greater than one, and the overall rank sum will exceed m.

Consequently, (28) will always be negative. This is because if $\sum_{i=1}^m z_i = k$ for any set of non-negative variables $\{z_i\}$, the minimal value of $\sum_{i=1}^m z_i^2$ occurs when $z_i = k/m$, $\forall i$. In our case, this implies that

$$\sum_{i=1}^m \left(\boldsymbol{b}_i^\top \widehat{\boldsymbol{\Sigma}}_{b_i}^{-1} \boldsymbol{b}_i \right)^2 \geq \sum_{i=1}^m (k/m)^2 > k > m. \tag{29}$$

Therefore we can add a small contribution of $\boldsymbol{b}_i \boldsymbol{b}_i^\top$ to each $\widehat{\boldsymbol{\Sigma}}_{b_i}$ and reduce the underlying cost function. Hence we cannot have a local minimum, except when $\boldsymbol{\Psi}$ is equal to some $\boldsymbol{\Psi}^*$ with $\mathrm{rank}[\boldsymbol{\Psi}^*] = 1$. Moreover, we may directly conclude that $\boldsymbol{x}^* = \overline{\boldsymbol{\Psi}}^* \boldsymbol{A}^\top (\boldsymbol{A} \overline{\boldsymbol{\Psi}}^* \boldsymbol{A}^\top)^\dagger \boldsymbol{b}$ is feasible and $\mathrm{rank}[\boldsymbol{X}^*] = 1$ with $\boldsymbol{x}^* = \mathrm{vec}[\boldsymbol{X}^*]$.

Regarding the last part of the theorem, we consider only f that are concave non-decreasing functions (this is the only reasonable choice for shrinking singular values to zero, and the more general case naturally follows anyway with additional effort, but minimal enlightenment). Without loss of generality we may also assume that $f(0) = 0$ and $f(1) = 1$; we can always apply an inconsequential translation and scaling such that these conditions hold.[9] Simple counter examples then demonstrate that $f(\epsilon)$ must be greater than some constant C independent of ϵ for all ϵ sufficiently small. To see this, note that we can always rescale elements of $\boldsymbol{A}$ such that a solution with rank greater than one is preferred unless this condition holds. However, such an f, which effectively must display infinite gradient at $f(0)$ to guarantee a global solution is always rank one, will then always display local minima for certain $\boldsymbol{A}$. This can easily be revealed through simple counter-examples. ■

C. Symmetrization Update Rules

These iterative update rules follow from alternative upper bounds tailored to the symmetric version of BARM. When both $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_c$ are fixed, $\boldsymbol{x}$ is updated via the posterior mean calculation

$$\begin{aligned} \widehat{\boldsymbol{x}} = \ & \mathrm{vec}\left[\widehat{\boldsymbol{X}}\right] = \frac{1}{2}\left(\overline{\boldsymbol{\Psi}}_r + \overline{\boldsymbol{\Psi}}_c\right) \boldsymbol{A}^\top \\ & \times \left[\lambda \boldsymbol{I} + \boldsymbol{A} \frac{1}{2}\left(\overline{\boldsymbol{\Psi}}_r + \overline{\boldsymbol{\Psi}}_c\right) \boldsymbol{A}^\top\right]^{-1} \boldsymbol{b}. \end{aligned} \tag{30}$$

where $\overline{\boldsymbol{\Psi}}_r = \boldsymbol{\Psi}_r \otimes \boldsymbol{I}$ and $\overline{\boldsymbol{\Psi}}_c = \boldsymbol{I} \otimes \boldsymbol{\Psi}_c$. Likewise we update $\nabla_{\Psi_r^{-1}}$ and $\nabla_{\Psi_c^{-1}}$ using

$$\nabla_{\Psi_r^{-1}} = \sum_{i=1}^m \boldsymbol{\Psi}_r - \boldsymbol{\Psi}_r \boldsymbol{A}_{\mathrm{ri}}^\top \left(\boldsymbol{A} \overline{\boldsymbol{\Psi}}_r \boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_{\mathrm{ri}} \boldsymbol{\Psi}_r, \tag{31}$$

$$\nabla_{\Psi_c^{-1}} = \sum_{i=1}^n \boldsymbol{\Psi}_c - \boldsymbol{\Psi}_c \boldsymbol{A}_{\mathrm{ci}}^\top \left(\boldsymbol{A} \overline{\boldsymbol{\Psi}}_c \boldsymbol{A}^\top + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{A}_{\mathrm{ci}} \boldsymbol{\Psi}_c, \tag{32}$$

[9]The log function is a limiting case, but what follows holds nonetheless.

where $\boldsymbol{A}_{\mathrm{ri}} \in \mathbb{R}^{p \times m}$ is defined such that $\boldsymbol{A} = [\boldsymbol{A}_{r1}^\top, \ldots, \boldsymbol{A}_{\mathrm{rm}}^\top]^\top$ and $\boldsymbol{A}_{\mathrm{ci}} \in \mathbb{R}^{p \times m}$ is defined such that $\boldsymbol{A} = [\boldsymbol{A}_{c1}, \ldots, \boldsymbol{A}_{\mathrm{cn}}]$. Finally given these values, with $\boldsymbol{X}$, $\nabla_{\Psi_r^{-1}}$ and $\nabla_{\Psi_c^{-1}}$ fixed, we can compute the optimal $\boldsymbol{\Psi}_r$ and $\boldsymbol{\Psi}_c$ in closed form by optimizing the relevant $\boldsymbol{\Psi}_r$- and $\boldsymbol{\Psi}_c$-dependent terms via

$$\boldsymbol{\Psi}_r^{\mathrm{opt}} = \frac{1}{n}\left[\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}} + \nabla_{\Psi_r^{-1}}\right], \tag{33}$$

$$\boldsymbol{\Psi}_c^{\mathrm{opt}} = \frac{1}{m}\left[\widehat{\boldsymbol{X}} \widehat{\boldsymbol{X}}^\top + \nabla_{\Psi_c^{-1}}\right]. \tag{34}$$

In practice the simple initialization $\boldsymbol{\Psi}_r = \boldsymbol{I}$ and $\boldsymbol{\Psi}_c = \boldsymbol{I}$ is sufficient for obtaining good performance.

REFERENCES

[1] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[2] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He, "Fast and accurate matrix completion via truncated nuclear norm regularization," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 35, no. 9, pp. 2117–2130, 2013.

[3] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 35, no. 1, pp. 171–184, 2013.

[4] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma, "Tilt: Transform invariant low-rank textures," *Int. J. Comput. Vis. (IJCV)*, vol. 99, no. 1, pp. 1–24, 2012.

[5] C. Lu, J. Tang, S. Yan, and Z. Lin, "Generalized nonconvex nonsmooth low-rank minimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[6] K. Mohan and M. Fazel, "Iterative reweighted algorithms for matrix rank minimization," *J. Mach. Learn. Res. (JMLR)*, vol. 13, no. 1, pp. 3441–3473, 2012.

[7] Z. Li, J. Liu, Y. Jiang, J. Tang, and H. Lu, "Low rank metric learning for social image retrieval," in *Pro. 20th ACM Int. Conf. Multimedia*, 2012, pp. 853–856.

[8] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc. B*, vol. 61, no. 3, pp. 611–622, 1999.

[9] B. Xin and D. Wipf, "Pushing the limits of affine rank minimization by adapting probabilistic pca," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 419–427.

[10] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[11] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos, "Sparse bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, 2012.

[12] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, 2011.

[13] D. Wipf, "Non-convex rank minimization via an empirical Bayesian approach," in *Proc. 28th Conf. Uncertainty Artif. Intell. (UAI)*, 2012.

[14] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[15] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res. (JMLR)*, vol. 1, pp. 211–244, 2001.

[16] D. P. Wipf, B. D. Rao, and S. Nagarajan, "Latent variable Bayesian models for promoting sparsity," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6236–6255, 2011.

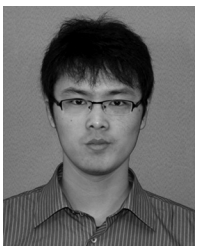
[17] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[18] E. J. Candes and X. Li, "Solving quadratic equations via phaselift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 14, no. 5, pp. 1017–1026, 2014.

[19] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, 2015.

[20] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*, Englewood Cliffs, NJ, USA: Prentice-Hall, 1969.

[21] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," *SIAM J. Scientif. Comput.*, vol. 35, no. 5, pp. S104–S125, 2013.
[22] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 937–945.
[23] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Found. Comput. Math.*, vol. 11, no. 2, pp. 183–210, 2011.
[24] R. H. Keshavan and S. Oh, *"A gradient descent algorithm on the Grassman manifold for matrix completion,"* 2009 DOI: arXiv Preprint arXiv:0910.5260.
[25] F. Léger, G. Yu, and G. Sapiro, *"Efficient matrix completion with Gaussian models,"* 2010 DOI: arXiv Preprint arXiv:1010.4050.
[26] B. Marlin, *Collaborative filtering: A machine learning perspective*, Ph.D. dissertation, Univ. of Toronto, Toronto, Canada ON, 2004.
[27] B. M. Marlin, "Modeling user rating profiles for collaborative filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003.
[28] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd ACM Int. Conf. Mach. Learn.*, 2005, pp. 713–719.
[29] S.-T. Park and D. M. Pennock, "Applying collaborative filtering techniques to movie search for better ranking and browsing," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2007, pp. 550–559.
[30] D. DeCoste, "Collaborative prediction using ensembles of maximum margin matrix factorizations," in *Proc. 23rd ACM Int. Conf. Mach. Learn.*, 2006, pp. 249–256.
[31] N. D. Lawrence and R. Urtasun, "Non-linear matrix factorization with gaussian processes," in *Proc. 26th Annu. ACM Int. Conf. Mach. Learn.*, 2009, pp. 601–608.
[32] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin, "Nonparametric Bayesian matrix completion," *Proc. IEEE SAM*, 2010.



Bo Xin (M'XX) received the B.S. degree in electronic engineering from Dalian University of Technology, China, in 2011. He is currently working toward the Ph.D. degree in computer science at Peking University, China. His research interests include optimization, machine learning and computer vision.

David Wipf (M'XX) received the B.S. degree with highest honors from the University of Virginia, and the Ph.D. degree from UC San Diego, where he was an NSF IGERT Fellow. Later he was an NIH Postdoctoral Fellow at UC San Francisco. Since 2011 he has been with Microsoft Research in Beijing. His research interests include Bayesian learning techniques applied to signal/image processing and computer vision. He is the recipient of several awards including the 2012 Signal Processing Society Best Paper Award, the Biomag 2008 Young Investigator Award, and the 2006 NIPS Outstanding Paper Award.

Yizhou Wang (M'XX) received his Ph.D. in computer science from University of California at Los Angeles (UCLA) in 2005. He was a Research Staff of the Palo Alto Research Center (Xerox-PARC) from 2005 to 2008. He is currently a Professor of the Computer Science Department at Peking University (PKU), China. His research interests include computer vision, statistical modeling and learning.

Wen Gao (F'XX) received M.S. degree in computer science from Harbin Institute of Technology in 1985, and Ph.D. degree in electronics engineering from the University of Tokyo in 1991. He was a Professor in computer science at Harbin Institute of Technology from 1991 to 1995 and a Professor in computer science at Institute of Computing Technology of Chinese Academy of Sciences from 1996 to 2005. He is currently a Professor at the School of Electronics Engineering and Computer Science, Peking University, China. He has been leading research efforts to develop systems and technologies for video coding, face recognition, sign language recognition and synthesis, and multimedia retrieval. He earned many awards, which include five national awards for his research achievements and activities. He did many services to academic society, such as general co-chair of IEEE ICME07, and the head of Chinese delegation to the Moving Picture Expert Group (MPEG) of International Standard Organization (ISO). Since 1997, he is also the chairman of the working group responsible for setting a national Audio Video coding Standard (AVS) for China. He published four books and over 500 technical articles in refereed journals and proceedings in the areas of signal processing, image and video communication, computer vision, multimodal interface, pattern recognition, and bioinformatics.

QUERIES

Q1. Author: For all conference paper references, provide page numbers if printed in proceeding or location of where conference was presented if not printed.

Q2. Author: Please provide a foward facing headshot for Dr. Wipf. Otherwise the bio will have to run without photo.

Q3. Author: Please provide initial year(s) of IEEE membership grade(s) for all authors.