Exploring Algorithmic Limits of Matrix Rank Minimization under Affine Constraints

Bo Xin† and David Wipf‡
†School of EECS, Peking University, Beijing, China
‡Visual Computing Group, Microsoft Research Asia, Beijing, China
jimxinbo@gmail.com davidwipf@gmail.com

Abstract

A wide variety of applications require recovering a matrix of minimal rank within an affine constraint set, with matrix completion a notable special case. Because the underlying 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 have been derived to elucidate when this replacement is likely to be successful, these conditions are highly restrictive and convex algorithms fail when the ambient rank is too high or when the constraint set is poorly structured. Non-convex alternatives fare somewhat better when carefully tuned; however, convergence to locally optimal solutions remains a continuing source of failure. Under this backdrop we derive a deceptively simple, parameter-free Bayesian-inspired algorithm that is capable, over a wide battery of empirical tests, of successful recovery even approaching the theoretical limit where the number of measurements is equal to 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 conditions remains evasive for non-convex algorithms, Bayesian or otherwise, we nonetheless quantify special cases where our algorithm is guaranteed to succeed while existing algorithms may fail. We conclude with a simple computer vision application involving image rectification.

1 Introduction

Recently there has been a surge of interest in finding minimum rank matrices subject to some problem-specific constraints often characterized as an affine set \([3, 2, 6, 7, 8, 9, 14]\). Mathematically this involves solving

\[
\min_X \text{rank}[X] \quad \text{s.t. } b = \mathcal{A}(X),
\]

where \(X \in \mathbb{R}^{n \times m}\) is the unknown matrix, \(b \in \mathbb{R}^p\) 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

\[
\min_X \text{rank}[X] \quad \text{s.t. } X_{ij} = (X_0)_{ij}, \quad (i, j) \in \Omega,
\]

where \(X_0\) is a low-rank matrix we would like to recover, but we are only able to observe elements from the set \(\Omega\) \([3, 6]\). Unfortunately however, both this special case and the general problem (1) are well-known to be an NP-hard, and the rank penalty itself is non-smooth. Consequently, a popular alternative is to instead compute

\[
\min_X \sum_i f(\sigma_i[X]) \quad \text{s.t. } b = \mathcal{A}(X),
\]
where \( \sigma_i[X] \) denotes the \( i \)-th singular value of \( X \) and \( f \) is usually a concave, non-decreasing function (or nearly so). In the special case where \( f(z) = 1[z \neq 0] \) (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 \([2,3,7,9]\). However, these guarantees typically only apply to a highly restrictive set of rank minimization problems, and in a practical setting non-convex algorithms can succeed in a much broader range of conditions \([6,8,9]\).

In Section 2 we will summarize state-of-the-art non-convex rank minimization algorithms that operate under affine constraints and point out some of their shortcomings. This will be followed in Section 3 by the derivation of an alternative approach using Bayesian probabilistic modeling techniques. Section 4 will then describe properties of global and local solutions as well as special cases where any minimizer is guaranteed to have optimal rank, illustrating the intrinsic underlying mechanism which leads to success over competing methods. Finally, Section 5 contains a wide variety of numerical comparisons that highlight the efficacy of our algorithm. In particular, to the best of our knowledge we demonstrate the only algorithm that is able to consistently recover an unknown low-rank matrix \( X \) right up to the theoretical limit where the number of degrees of freedom equals the number of measurements. Surprisingly, this is even possible when the linear operator \( \mathcal{A} \) may be ill-conditioned and long after any convex guarantees break down.

2 Related Work

Here we focus on a few of the latest and most effective rank minimization algorithms, all developed within the last year or two and evaluated favorably against the state-of-the-art. In the non-convex regime, effective optimization strategies attempt to at least locally minimize (3), often exceeding the performance of the convex nuclear norm. For example, [9] derives a family of iterative reweighted least squares (IRLS) algorithms applied to \( f(z) = (z^2 + \gamma)^{q/2} \) with \( q, \gamma > 0 \) as tuning parameters. A related penalty also considered, which coincides with the limit as \( q \to 0 \) (up to an inconsequential scaling and translation), is \( f(z) = \log(z^2 + \gamma) \). This case also maintains an intimate connection with rank given that \( \log z = \lim_{q \to 0} q^{-1}(z^q - 1) \equiv I[z \neq 0] \). Consequently, when \( \gamma \) is small, \( \sum_i \log(\sigma_i[X]^2 + \gamma) \) behaves much like the rank, albeit with nonzero gradients away from zero.

The IRLS0 algorithm from [9] represents the best-performing special case of the above, where \( \sum_i \log(\sigma_i[X]^2 + \gamma) \) is minimized using a homotopy continuation scheme merged with IRLS. Here a fixed \( \gamma \) is replaced with a decreasing sequence \( \{\gamma^k\} \), the rationale being that when \( \gamma_k \) is large, the cost function is relatively smooth and devoid of local minima. As the iterations \( k \) progress, \( \gamma^k \) 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 however is that it requires a pre-defined heuristic for reducing \( \gamma^k \), and this schedule may be problem specific. Moreover, there is no guarantee that a global solution will ever be found.

In a related vein, [8] 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 [6]. While some competitive results for image inpainting via matrix completion are shown,
practice the proposed algorithm has many parameters to be tuned via cross-validation. Moreover, recent comparisons contained in [8] show that default settings perform relatively poorly.

Finally, from an algorithmic standpoint, previous work has applied Bayesian formalisms to rank minimization problems, although not specifically within an affine constraint set. For example, [1][4][11] derive robust PCA algorithms built upon the linear summation of a rank penalty and an element-wise sparsity penalty. But none of these algorithms have been augmented and rigorously analyzed in the context of rank minimization with affine constraints.

3 Alternative Algorithm Derivation

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

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

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 \( \nu_i \Psi \) on each column of \( X \), denoted \( x_i \), for all \( i = 1, \ldots, m \). This produces the aggregate prior on \( X \) given by

\[ p(X; \Psi, \nu) = \prod_i \mathcal{N}(x_i; 0, \nu_i \Psi) \propto \exp \left[ x^\top \Psi^{-1} x \right], \]

where \( \Psi \in \mathbb{R}^{n \times n} \) is a positive semi-definite symmetric matrix, \( \nu = [\nu_1, \ldots, \nu_m]^\top \) is a non-negative vector, \( x = \text{vec}[X] \) (column-wise vectorization), and \( \Psi = \text{diag}[^\nu] \otimes \Psi \), with \( \otimes \) denoting the Kronecker product. Given both likelihood and prior are Gaussian, the posterior \( p(X|b; \mathcal{A}, \lambda) \) is also Gaussian, with mean given by an \( X \) such that

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

Here \( A \in \mathbb{R}^{p \times nm} \) is a matrix defining the linear operator \( \mathcal{A} \) such that \( b = Ax \) reproduces the feasible region in [1]. From this expression it is clear that, if \( \Psi \) represents a low-rank covariance matrix, then each column of \( \hat{X} \) 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 \( \Psi \) and the associated scaling vector \( \nu \).

A common Bayesian strategy in this regard is to marginalize over \( X \) and then maximize the resulting likelihood function with respect to \( \Psi \) and \( \nu \) [10][11][12]. This involves solving

\[ \max_{\Psi \in H^+, \nu \geq 0} \int p(b|X; \mathcal{A}, \lambda)p(X; \Psi, \nu) dX, \]

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

\[ \mathcal{L}(\Psi, \nu) = b^\top \Sigma_b b + \log |\Sigma_b|, \quad \Sigma_b = A\hat{\Psi} A^\top + \lambda I, \quad \Psi = \text{diag}[\nu] \otimes \Psi, \]

where \( \Sigma_b \) can be viewed as the covariance of \( b \) given \( \Psi \) and \( \nu \).

Minimizing [8] is a non-convex optimization problem, and we employ standard upper bounds for this purpose leading to an EM-like algorithm. In particular, we compute separate bounds, parameterized by auxiliary variables, for both the first and second terms of \( \mathcal{L}(\Psi, \nu) \). 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 \( \mathcal{L}(\Psi) \triangleq \mathcal{L}(\Psi, \nu = 1) \), meaning all elements of \( \nu \) are fixed at one (and such is the case for all experiments reported herein, although we are currently exploring situations where this added generality is especially helpful).

Based on [12], for the first term in [8] we have

\[ b^\top \Sigma_b b \leq \frac{1}{\lambda} \| b - Ax \|_2^2 + x^\top \Psi^{-1} x \]
with equality whenever $x$ satisfies (6). For the second term we use
\[
\log |\Sigma_b| = m \log |\Psi| + \log |\lambda A^\top A + \overline{\Psi}^{-1}| \leq m \log |\Psi| + \text{tr} [\overline{\Psi}^{-1} \nabla \Psi^{-1}] + C, \tag{10}
\]
where because $\log |\lambda A^\top A + \overline{\Psi}^{-1}|$ is concave with respect to $\Psi^{-1}$, we can upper bound it using a first-order approximation with a bias term $C$ that is independent of $\Psi$. Equality is obtained when the gradient satisfies
\[
\nabla \Psi^{-1} = \sum_{i=1}^m \Psi - \Psi A_i^\top \left( A \Psi A^\top + \lambda I \right)^{-1} A_i \Psi, \tag{11}
\]
where $A_i \in \mathbb{R}^{p \times n}$ is defined such that $A = [A_1, \ldots, A_m]$. Finally given the upper bounds from (9) and (10) with $X$ and $\nabla \Psi^{-1}$ fixed, we can compute the optimal $\Psi$ in closed form by optimizing the relevant $\Psi$-dependent terms via
\[
\Psi^{opt} = \arg\min_X \left[ \Psi^{-1} \left( XX^\top + \nabla \Psi^{-1} \right) \right] + m \log |\Psi| = \frac{1}{m} \left[ XX^\top + \nabla \Psi^{-1} \right]. \tag{12}
\]
By iteratively computing (6), (11), and (12), we can then obtain an estimate for $\Psi$, and more importantly, a corresponding estimate for $X$ given by (6) at convergence. We refer to this seemingly unremarkable algorithm as BARM for Bayesian Affine Rank Minimization, and the next section will describe in detail why it is particularly well-suited for solving problems such as (1).

4 Properties of BARM

As discussed in Section 2 one nice property of the $\sum_i \log |\sigma_i[X]|$ penalty employed (approximately) by IRLS0 [9] 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 $\sum_i \log |\sigma_i[X]|^2 + \gamma$ when $\gamma$ 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 $X^*$ is an optimal feasible solution, a block-diagonal rescaling of $A$ nevertheless leads to an equivalent rescaling of the optimum (without the need for solving an additional optimization problem using the new $A$). This is very much unlike the nuclear norm or other non-convex surrogates that penalize the singular values of $X$ 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 $\sum \log |\sigma_i[X]|^2 + \gamma$ with $\gamma \to 0$. Before presenting these results, we define $\text{spark}[A]$ as the smallest number of linearly dependent columns in a matrix $A$ [3]. All proofs are deferred to the appendix.

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

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

**Remarks:** The assumption $r = \text{rank}[X^*] < p/m$ in Lemma 1 is completely unrestricted, especially given that a unique, minimal-rank solution is only theoretically possible by any algorithm if $p \geq (m + m)r - r^2$, which is much more restrictive than $p > rm$. Hence the bound we require is well above that required for uniqueness anyway. Likewise the spark assumption will be satisfied for any $A$ 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 $\sum_i f(\sigma_i[X])$ 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 made in Section 2).
While these results are certainly a useful starting point, the real advantage of adopting the BARM cost function is that locally minimizing solutions are exceedingly rare, largely as a consequence of the marginalization process in (7), and in some cases provably so. A specialized example of this smoothing can be quantified in the following scenario.

Suppose A is now block diagonal, with diagonal blocks Aᵢ such that bᵢ = Aᵢxᵢ producing the aggregate observation vector b = [b₁, ..., bₘᵀ]. While somewhat restricted, this situation nonetheless includes many important special cases, including canonical matrix completion and generalized matrix completion where elements of Z = WXX₀ are observed instead of X₀ directly.

Theorem 1. Let b = A vec[X], where A is block diagonal, with blocks Aᵢ ∈ ℝᵖᵢ×ₙᵢ. Moreover, assume pᵢ > 1 for all i and that ∩ᵢ null[Aᵢ] = ⊥. Then if min_X rank[X] = 1 in the feasible region, then any minimizer {Ψ*, ν*} of (2) (global or local) in the limit λ → 0 is such that x* = ΨᵀAᵀ (AΨᵀAᵀ)⁻¹ b is feasible and rank[X*] = 1 with vec[X*] = 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 A, the proposed cost function has no minima that are not global minima. (The minor technical condition regarding nullspace intersections merely ensures that high-rank components cannot simultaneously “hide” in the nullspace of every measurement matrix Aᵢ; the actual 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. Moreover, such a guarantee is not possible with any other penalty function of the standard form ∑ᵢ f(σᵢ[Xᵢ]), which is the typical recipe for rank minimization algorithms, convex or not. Additionally, if a unique rank-one solution exists to (1), then the unique minimizing solution to (8) will produce this X via (6). Crucially, this will occur even when the minimal number of measurements p = n + m − 1 are available, unlike any other algorithm we are aware of that is blind to the true underlying rank. And the underlying intuition, that local minima are smoothed away, nonetheless carries over to situations where the rank is greater than one.

Visualization: To further explore this smoothing effect and complement Theorem 1, it helps to visualize rank penalty functions restricted to the feasible region. While the BARM algorithm involves minimizing (8), its implicit penalty function on 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 X* = VΨ*, where VΨ* is the true minimum rank solution, V ∈ null[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 V = X¹ − X*, where X¹ is the feasible solution with minimum nuclear norm; however, random selections from null[A] also show similar characteristics.

Figure 1 provides a simple example of this process. A is generated randomly with all zeros and a single randomly placed ‘1’ in each row leading to a canonical matrix completion problem. X* ∈ ℝ¹⁵×⁵ is randomly generated as X* = uvᵀ, where u and v are iid N(0, 1) vectors, and so X* is rank one. Finally, p = 10 elements are observed, and therefore A has 10 rows and 5 × 5 = 25 columns. α is varied from −5 to 5 and the values of the nuclear norm, ∑ᵢ log|σᵢ[Xᵢ]|² + γ|, 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 α = 0; hence the convex cost function fails for this problem. Likewise, the ∑ᵢ log|σᵢ[Xᵢ]|² + γ| penalty used by IRLS0 displays an incorrect global minimum when the tuning parameter γ is large. In contrast, when λ is small, while the global minimum may now be correct, spurious local ditches have appeared in the cost function. Therefore, any success of the IRLS0 algorithm depends heavily on a carefully balanced decaying sequence of γ values, with the

1It 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 X to the number of observations p. Consequently, estimating X even with r small can be challenging when p is also small, meaning A is highly overcomplete.

2Technically 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.
Figure 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 $\sum \log |\sigma_i| [X]^2 + \gamma|$ penalty exhibits the correct minimum when $\gamma$ is sufficiently small, it also contains spurious minima. Only BARM smoothes away local minimum while simultaneously retaining the correct global optima.

hope that initial iterations can steer the trajectory towards a desirable basin of attraction where local minima are less problematic. In this regard, the advantage of BARM is that it is parameter free and yet still retains the correct global minimum, often without additional spurious local minima.

Convergence: Regarding the BARM algorithm itself, the updates generated by (6), (11), and (12) are guaranteed to reduce or leave unchanged $\mathcal{L}(\Psi)$ at each iteration. However, this is not technically sufficient to guarantee convergence to a stationary point of the underlying cost function unless the additional conditions of Zangwill’s Global Convergence Theorem are satisfied [13]. However, provided we add a small regularization factor $\gamma \text{tr}(\Psi^{-1})$, with $\gamma > 0$, then it can be shown that any cluster point of the resulting sequence of iterations $\{\Psi^k\}$ 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 parameters values $S$ such that for any $\Psi^* \in S$, $\mathcal{L}(\Psi^*) + \gamma \text{tr}(\Psi^*^{-1})$ is a stationary point. Finally, we should mention that this extra $\gamma$ factor is akin to the homotopy continuation regularizer used by the IRLS0 algorithm [9] as discussed in Section 2. However, whereas IRLS0 requires a carefully-chosen, decreasing sequence $\{\gamma_k\}$ with $\gamma_k > 0$ both to prove convergence and avoid local minimum (and without this factor the algorithm performs very poorly in practice), for BARM a small, fixed factor only need been included as a technical necessity for proving formal convergence; in practice it can be fixed to exactly zero.

Symmetrization Improvements: Despite the promising theoretical attributes of BARM, there remains one potential artifact of its Bayesian 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 $X$ or $X^\top$. 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 $X$. Consequently the underlying cost function is not identical when derived with respect to $X$ or $X^\top$, a difference which will depend on $A$. 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 $x = \text{vec}[X]$ with a covariance formed using a block-wise averaging of covariances defined over rows and columns, denoted $\Sigma_r$ and $\Sigma_c$, respectively. The overall covariance is then given by $\Psi = 1/2 (\Sigma_r \otimes I + I \otimes \Sigma_c)$. The estimation process then proceeds in a similar fashion as before with slight modifications to adapt for this merger.

5 Experimental Validation

This section compares BARM with existing state-of-the-art affine rank minimization algorithms. For BARM, in all cases we simply used $\lambda = 10^{-10}$ (effectively zero), and hence no tuning parameters are required. Likewise, nuclear norm minimization [5] 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 based the decreasing $\gamma_k$ sequence on suggestions from [9]. For the IRNN algorithm of [18], we did not have access to code nor specific details of how various parameters were set. Therefore we
only report results directly from [8] where available. Note that both [8] and [9] show superior results to a number of other algorithms; however, we do not directly compare with any these others given that they are likely no longer state-of-the-art and may clutter the presentation.

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 [8], where a rank $A$ matrix is generated as: (i) an iid $X$ vector. The latter is meant to explore less-than-ideal conditions where the linear operator displays correlations and may be somewhat ill-conditioned. Figure 3 displays aggregate results when $X_0$, in this case $2 \times 150 \times 44 - 44^2 = 11264$. Note that FoS values were reported in [8] over a wide range of non-convex IRNN algorithms. The green curve represents the best performing candidate from this pool; REL values were unavailable.

General $A$: Next we consider the more challenging problem involving arbitrary affine constraints. $X_0$ is generated in the same way as before. We then consider two types of linear mappings where $A$ is generated as: (i) an iid $N(0, 1)$, $p \times n^2$ matrix, and (ii) $\sum_{i=1}^{p} i^{-1/2} u_i v_i^\top$, where $u_i \in \mathbb{R}^p$ and $v_i \in \mathbb{R}^{n^2}$ are iid $N(0, 1)$ vectors. The latter is meant to explore less-than-ideal conditions where the linear operator displays correlations and may be somewhat ill-conditioned. Figure 3 displays aggregate results when $X_0$ is $50 \times 50$ and $100 \times 100$. In both cases $p = 1000$ observations are used, therefore the corresponding measurement matrices $A$ are $1000 \times 2500$ and $1000 \times 10000$ 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 $A$ operator is no longer an ideal Gaussian. Note that the d.o.f. for the $50 \times 50$ and $100 \times 100$ cases when $r = 11$ and $r = 5$ are 979 and 975 respectively. Hence we have even tested reducing $p$ all the way down to 979 and 975 and still BARM was able to sometimes succeed (see appendix, as well as further examination of failure cases). In general, we have explored a wide range of empirical conditions too lengthy to report here, and it is only very rarely, and always near the theoretical boundary, where BARM occasionally may not succeed.

Application Example: In [14], the TILT algorithm is derived for rectifying images containing low-rank textures that have been transformed using an unknown operator $\tau$ from some group (e.g., a homography). For a given observed image $Y$, the basic idea is to construct a first-order Taylor series approximation around the current rectified image estimate $X$ and solve

$$\min_X \text{rank}[X] \quad \text{s.t.} \quad X = Y + \sum_i J_i(\hat{X}) \delta_i,$$

where $J_i(\hat{X})$ is the Jacobian matrix with respect to $X$ of the $i$-th parameter $\tau_i$ describing the transformation. Optimization over $\delta$ the vector of first-order differences can be accomplished in closed form by projecting both sides of the constraint to the orthogonal complement of the span of all $J_i(\hat{X})$. The resulting problem then reduces exactly to (1). Once $X$ is computed this way, we then update each $J_i(\hat{X})$ and repeat until convergence.
While the original TILT algorithm substitutes the nuclear norm for $\text{rank}[X]$, we embedded the BARM algorithm into the posted TILT source code [14] 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). Figure 4 displays results on one easy example, where the number of observations $p$ is large, and another more difficult problem where the number of observations is small. While both algorithms succeed on the easy case, when the observations are constrained by a small image window, only BARM is successful.

6 Conclusion

This paper explores a conceptually simple, parameter-free algorithm for matrix rank minimization under affine constraints that is capable of successful recovery empirically observed to approach the theoretical limit over a broad class of experimental settings (including many not shown here), unlike existing algorithms. Regarding computational complexity, for general $A$ (the special case of matrix completion is much easier), BARM scales linearly in the elements of $X$ and quadratically in the number of observations. In our experiments, for relatively easy problems on the order of 10 to 20 iterations are required, while for difficult recovery problems near the boundary this may increase by a factor of 10 or so. This is somewhat expected though since as we near the theoretical limit, $A$ becomes highly overcomplete, and candidate solutions become much more difficult to differentiate.
A Proof Sketch of Lemmas 1 and 2

Regarding Lemma 1, this result mirrors related ideas from [12] in the context of Bayesian compressive sensing. While a more rigorous presentation is possible, we defer such treatment to a later journal publication, and here just describe the basic outline for brevity. At any candidate solution define \(\Psi W\) such that \(\Sigma_b = WW^T\). To be a minimizer, global or local, it must be that \(b \in \text{span}[W]\). If this were not the case, then \(L(\Psi, \nu)\) would diverge to infinity because the \(\log |\Sigma_b|\) term is not sufficiently strong counterbalance, in the same way that \(\min \frac{1}{z} + \log z\) equals one (finite). Consequently, the only way to minimize the cost in the limit as \(\lambda \to 0\) is to consider low-rank solutions within this constraint set, and it is equivalent to requiring that \(b^T \Sigma_b^{-1} b \leq C\) for some constant independent of \(\lambda\) (which ultimately corresponds with maintaining \(A(X) = b\) in the limit as well).

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

\[
\log |\Sigma_b| = \sum_{i=1}^{p} \log \left( \sigma_i[A \Psi A^T] + \lambda I \right),
\]

where \(\sigma_i[\cdot]\) denotes the \(i\)-th singular value of a matrix. While the maximum rank of \(A \Psi A^T\) is obviously \(p\), if \(\text{rank}[\Psi] < p/m\) and \(\text{spark}[A] = p + 1\) (maximal spark) as stipulated in the lemma statement, then \(r \equiv \text{rank}[A \Psi A^T] = m \text{rank}[\Psi]\) and \((15)\) becomes

\[
\log |\Sigma_b| = \sum_{i=1}^{mr} \log \left( \sigma_i[A \Psi A^T] + \lambda I \right) + (p - mr) \log \lambda.
\]

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

Consequently, in the limit where \(\lambda \to 0\) (with the limit being taken outside of the minimization), \((15)\) effectively scales as \((p - mr) \log \lambda\), and hence the overall cost is minimized when \(\Psi\) has minimal rank. This in turn ensures that the corresponding \(X\) will also have minimal rank, completing the proof sketch for Lemma 1.

Finally, Lemma 2 follows directly from the structure of the \(L(\Psi, \nu)\) cost function via simple reparameterizations.

B Proof Sketch of Theorem 1

To begin we assume that \(b_i \neq 0, \forall i\), where \(b_i\) denotes the sub-vector of \(b\) such that \(b_i = A_i x_i\). If this were not the case we can always collapse \(X\) 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

\[
L(\Psi, \nu) = \sum_{i=1}^{m} b_i^T \left( \nu_i A_i \Psi A_i^T \right)^{-1} b_i + \log \left| \nu_i A_i \Psi A_i^T \right|.
\]

If there exists a feasible rank one solution to \(b = A \text{vec}[X]\), then there exists a set of \(\Psi' = \nu_i \hat{\Psi}\) such that \(b_i b_i^T = A_i \hat{\Psi}_i A_i^T\) for all \(i\). To see this, note that \(b_i b_i^T = A_i x_i x_i^T A_i^T\). Because \(\text{rank}[X] = 1\), it also follows that \(b_i b_i^T = \alpha_i A_i X X^T A_i^T\), where \(\alpha_i = ||x_i|| \text{||} \text{X} \text{||}\). Therefore \(\Psi'_i = \nu_i X \text{||} A_i \text{||}\) achieves the desired result with \(\nu_i = \alpha_i\).

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

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

\[
\mathcal{L}(\lambda_1, \lambda_2) = \sum_{i=1}^{m} b_i^\top \left( \lambda_1 \Sigma_{b_i} + \lambda_2 b_i b_i^\top \right)^{-1} b_i + \log \left| \lambda_1 \Sigma_{b_i} + \lambda_2 b_i b_i^\top \right| .
\] (17)

This is because \( b_i b_i^\top \) represents a valid set of basis vectors for updating the covariance per the construction above involving \( \Psi_i \). First consider optimization over \( \lambda_1 \). If \( \lambda_1 = 1 \) is a local minimum, then by taking gradients and equating to zero, we require that

\[
\sum_{i=1}^{m} b_i^\top \Sigma_{b_i}^{-1} b_i = \sum_{i=1}^{m} \text{rank}[\hat{\Sigma}_{b_i}].
\] (18)

Likewise, taking the gradient with respect to \( \lambda_2 \) we obtain

\[
\frac{\partial \mathcal{L}(\lambda_1, \lambda_2)}{\partial \lambda_2} \bigg|_{\lambda_1=1, \lambda_2=0} = \sum_{i=1}^{m} b_i^\top \Sigma_{b_i}^{-1} b_i - \sum_{i=1}^{m} \left( b_i^\top \Sigma_{b_i}^{-1} b_i \right)^2 .
\] (19)

The nullspace condition (a very mild assumption) ensures that \( \sum_{i=1}^{m} \text{rank}[\hat{\Sigma}_{b_i}] = k \) for some \( k > m \) when \( \text{rank}[\Psi] > 1 \). To see this, observe that to achieve \( \sum_{i=1}^{m} \text{rank}[\hat{\Sigma}_{b_i}] = m \) when \( \text{rank}[\Psi] > 1 \) requires that \( \Psi = uu^\top + WW^\top \) where \( u \) is a vector and \( W \) is a matrix (or vector) with columns in \( \text{null}[A_i], \forall i \). If any such \( W \) is not in this nullspace for some \( i \), then given that \( p_i > 1 \), the associated \( A_i \Sigma A_i^\top \) will have rank greater than one, and the overall rank sum will exceed \( m \).

Consequently, (19) 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

\[
\sum_{i=1}^{m} \left( b_i^\top \Sigma_{b_i}^{-1} b_i \right)^2 \geq \sum_{i=1}^{m} (k/m)^2 > k > m.
\] (20)

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

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 minimal effort). 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\(^\ddagger\). Simple counter examples then demonstrate that \( f(\epsilon) \) must be greater than some constant \( C \) independent of \( \epsilon \) for all \( \epsilon \) sufficiently small. To see this, note that we can always rescale elements of \( A \) 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 \( A \). This can easily be revealed through simple counter-examples. □

### C Failure Case Analysis

With the number of observations fixed at \( p = 1000 \) and a general measurement matrix \( A \), in Section 5 of our main paper we 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 this 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 \( p = 979 \), while for \( \{n = 100, r = 11\} \) this occurs at \( p = 975 \).

\(^\ddagger\)The log function is a limiting case, but what follows holds nonetheless.
We examined the BARM algorithm under these conditions with 10 additional trials 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 Figure 5 we have averaged the singular values of all the failures. Here 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 rank. We therefore speculate that right at the theoretical limit, when \(A\) is maximally overcomplete \((p \times n^2 = 979 \times 2500\) or \(975 \times 10000\) for the two problem sizes), there exists multiple feasible matrices with rank and singular value spectrum nearly indistinguishable from the optimal solution. Note that when the other algorithms we tested failed, the failure is much more dramatic and a clear cut-off at the correct rank is not so apparent.

References

[1] S Derin Babacan, Martin Luessi, Rafael Molina, and Aggelos K Katsaggelos. Sparse bayesian methods for low-rank matrix estimation. Signal Processing, IEEE Transactions on, 60(8):3964–3977, 2012.
[2] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? Journal of the ACM (JACM), 58(3):11, 2011.
[3] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. Foundations of Computational Mathematics, 9(6):717–772, 2009.
[4] Xinghao Ding, Lihan He, and Lawrence Carin. Bayesian robust principal component analysis. Image Processing, IEEE Transactions on, 20(12):3419–3430, 2011.
[5] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. Proceedings of the National Academy of Sciences, 100(5):2197–2202, 2003.
[6] Yao Hu, Debing Zhang, Jieping Ye, Xuelong Li, and Xiaofei He. Fast and accurate matrix completion via truncated nuclear norm regularization. Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on, 35(9):2117–2130, 2013.
[7] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on, 35(1):171–184, 2013.
[8] Canyi Lu, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference on. IEEE, 2014.
[9] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. The Journal of Machine Learning Research (JMLR), 13(1):3441–3473, 2012.
[10] Michael E Tipping. Sparse bayesian learning and the relevance vector machine. The Journal of Machine Learning Research (JMLR), 1:211–244, 2001.
[11] David Wipf. Non-convex rank minimization via an empirical bayesian approach. In Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI), 2012.

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

[13] Willard I Zangwill. Nonlinear programming: a unified approach. Prentice Hall, 1969.

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