# A Pseudo-Bayesian Algorithm for Robust PCA

**Authored by:**

David Wipf
Tae-Hyun Oh
Yasuyuki Matsushita
In Kweon

**Abstract**

Commonly used in many applications, robust PCA represents an algorithmic attempt to reduce the sensitivity of classical PCA to outliers. The basic idea is to learn a decomposition of some data matrix of interest into low rank and sparse components, the latter representing unwanted outliers. Although the resulting problem is typically NP-hard, convex relaxations provide a computationally-expedient alternative with theoretical support. However, in practical regimes performance guarantees break down and a variety of non-convex alternatives, including Bayesian-inspired models, have been proposed to boost estimation quality. Unfortunately though, without additional a priori knowledge none of these methods can significantly expand the critical operational range such that exact principal subspace recovery is possible. Into this mix we propose a novel pseudo-Bayesian algorithm that explicitly compensates for design weaknesses in many existing non-convex approaches leading to state-of-the-art performance with a sound analytical foundation.

## 1 Paper Body

It is now well-established that principal component analysis (PCA) is quite sensitive to outliers, with even a single corrupted data element carrying the potential of grossly biasing the recovered principal subspace. This is particularly true in many relevant applications that rely heavily on lowdimensional representations [8, 13, 27, 33, 22]. Mathematically, such outliers can be described by the measurement model Y = Z + E, where Y ? Rn?m is an observed data matrix, Z = AB¿ is a low-rank component with principal subspace equal to span[A], and E is a matrix of unknown sparse corruptions with arbitrary amplitudes. Ideally, we would like to remove the effects of E, which would then allow regular PCA to be applied to Z for obtaining principal components devoid of unwanted bias. For this purpose, robust PCA (RPCA) algorithms have recently been motivated by the optimization problem minZ,E max(n, m) ? rank[Z] + kEk0 s.t. Y = Z

+ E,
(1)

where k ? k0 denotes the ‘0 matrix norm (meaning the number of nonzero matrix elements) and the max(n, m) multiplier ensures that both rank and sparsity terms scale between 0 and nm, reflecting a priori agnosticism about their relative contributions to Y. The basic idea is that if {Z? , E? } minimizes (1), then Z? is likely to represent the original uncorrupted data. As a point of reference, if we somehow knew a priori which elements of E were zero (i.e., no gross corruptions), then (1) could be effectively reduced to the much simpler matrix completion (MC) problem [5] minZ rank[Z] s.t. yij = zij , ?(i, j) ? ?, (2) where ? denotes the set of indices corresponding with zero-valued elements in E. A major challenge with RPCA is that an accurate estimate of the support set ? can be elusive. ?

This work was done while the first author was an intern at Microsoft Research, Beijing. The first and third authors were supported by the NRF of Korea grant funded by the Korea government, MSIP (No. 2010-0028680). The second author was partly supported by JSPS KAKENHI Grant Number JP16H01732.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Unfortunately, solving (1) is non-convex, discontinuous, and NP-hard in general. Therefore, the convex surrogate referred to as principal component pursuit (PCP) p minZ,E max(n, m) ? kZk? + kEk1 s.t. Y = Z + E (3) is often adopted, where k ? k? denotes the nuclear norm and k ? k1 is the ‘1 matrix norm. These represent the tightest convex relaxations of the rank and ‘0 norm functions respectively. Several theoretical results quantify technical conditions whereby the solutions of (1) and (3) are actually equivalent [4, 6]. However, these conditions are highly restrictive and do not provably hold in practical situations of interest such as face clustering [10], motion segmentation [10], high dynamic range imaging [22] or background subtraction [4]. Moreover, both the nuclear and ‘1 norms are sensitive to data variances, often over-shrinking large singular values of Z or coefficients in E [11]. All of this motivates stronger approaches to approximating (1). In Section 2 we review existing alternatives, including both non-convex and probabilistic approaches; however, we argue that none of these can significantly outperform PCP in terms of principal subspace recovery in important, representative experimental settings devoid of prior knowledge (e.g., true signal distributions, outlier locations, rank, etc.). We then derive a new pseudo-Bayesian algorithm in Section 3 that has been tailored to conform with principled overarching design criteria. By ?pseudo?, we mean an algorithm inspired by Bayesian modeling conventions, but with special modifications that deviate from the original probabilistic script for reasons related to estimation quality and computational efficiency. Next, Section 4 examines relevant theoretical properties, explicitly accounting for all approximations involved, while Section 5 provides empirical validations. Proofs and other technical details are deferred to [23]. Our high-level contributions can be summarized as follows: - We derive a new pseudo-Bayesian RPCA algorithm with efficient ADMM subroutine. - While provable recovery guarantees are absent for non-convex RPCA

algorithms, we nonetheless quantify how our pseudo-Bayesian design choices lead to a desirable energy landscape. In particular, we show that although any outlier support pattern will represent an inescapable local minima of (1) (or a broad class of functions that mimic (1)), our proposal can simultaneously retain the correct global optimum while eradicating at least some of the suboptimal minima associated with incorrect outlier location estimates. - We empirically demonstrate improved performance over state-of-the-art algorithms (including PCP) in terms of standard phase transition plots with a dramatically expanded success region. Quite surprisingly, our algorithm can even outperform convex matrix completion (MC) despite the fact that the latter is provided with perfect knowledge of which entries are not corrupted, suggesting that robust outlier support pattern estimation is indeed directly facilitated by our model.

2

Recent Work

The vast majority of algorithms for solving (1) either implicitly or explicitly attempt to solve a problem of the form P minZ,E f1 (Z) + i,j f2 (eij ) s.t. Y = Z + E, (4) where f1 and f2 are penalty functions that favor minimal rank and sparsity respectively. When f1 is the nuclear norm (scaled appropriately) and f2 (e)=—e—, then (4) reduces to (3). Methods differ however by replacing f1 and f2 with non-convex alternatives, such as generalized Huber functions [7] or Schatten ‘p quasi-norms with p ¡ 1 [18, 19]. When applied to the singular values of Z and elements of E respectively, these selections enact stronger enforcement of minimal rank and sparsity. If prior knowledge of the true rank of Z is available, a truncated nuclear norm approach (TNN-RPCA) has also been proposed [24]. Further divergences follow from the spectrum of optimization schemes applied to different objectives, such as the alternating directions method of multipliers (ADMM) algorithm [3] or iteratively reweighted least squares (IRLS) [18]. With all of these methods, we may consider relaxing the strict equality constraint to the regularized form P minZ,E ?1 kY ? Z ? Ek2F + f1 (Z) + i,j f2 (eij ), (5) where ? ¿ 0 is a trade-off parameter. This has inspired a number of competing Bayesian formulations, which typically proceed as follows. Let

1 p(Y—Z, E) ? exp ? 2? kY ? Z ? Ek2F (6) 2

define a likelihood function, where ? represents a non-negative variance parameter assumed to be known.2 Hierarchical prior distributions are then assigned to Z and E to encourage minimal rank and strong sparsity, respectively. For the latter, the most common choice is the Gaussian scale-mixture (GSM) defined hierarchically by h 2 i Q e ?1 1?a ) ? ?ij exp[ ??b ], p(E—?) = i,j p(eij —?ij ), p(eij —?ij ) ? exp ? 2?ijij , with hyper prior p(?ij ij (7) where ? is a matrix of non-negative variances and a, b?0 are fixed parameters. Note that when these values are small, the resulting distribution over each eij (obtained by marginalizing over the respective ?ij ) is heavy-tailed with a sharp peak at zero, the defining characteristics of sparse priors. For the prior on Z, Bayesian methods have somewhat broader distinctions. In particular, a number of methods explicitly assume that Z=AB¿ and specify GSM priors on A and B [1, 9, 15, 30]. For ex ample, variational Bayesian RPCA (VB-RPCA) [1] assumes p(A—?)?exp ?tr A diag[?]?1 A¿ , where ? is a non-negative variance vector. An

equivalent Q prior is used for p(B—?) with a shared value of ?. This model also applies the prior p(?) = i p(?i ) with p(?i ) defined for consistency ?1 with p(?ij ) in (7). Low rank solutions are favored via the same mechanism as described above for sparsity, but only the sparse variance prior is applied to columns of A and B, effectively pruning them from the model if the associated ?i is small. Given the above, the joint distribution is p(Y, A, B, E, ?, ?) = p(Y—A, B, E)p(E—?)p(A—?)p(B—?)p(?)p(?).

(8)

Full Bayesian inference with this is intractable, hence a common variational Bayesian (VB) meanfield approximation is applied [1, 2]. The basic idea is to obtain a tractable approximate factorial posterior distribution by solving minq(?) KL [q(?)——p(A, B, E, ?, ?—Y)] ,

(9)

where q(?) , q(A)q(B)q(E)q(?)q(?), each q represents an arbitrary probability distribution, and KL[?——?] denotes the Kullback-Leibler divergence between two distributions. This can be accomplished via coordinate descent minimization over each respective q distribution while holding the others fixed. Final estimates of Z and E are obtained by the means of q(A), q(B), and q(E) upon convergence. A related hierarchical model is used in [9, 30], but MCMC sampling techniques are used for full Bayesian inference RPCA (FB-RPCA) at the expense of considerable computational complexity and multiple tuning parameters. An alternative empirical Bayesian algorithm (EB-RPCA) is described in [31]. In addition to the likelihood function (6) and prior from (7), this method assumes a direct Gaussian prior on Z given by

p(Z—?) ? exp ? 12 tr Z¿ ??1 Z , (10) where ? is a symmetric and positive definite matrix.3 Inference is accomplished via an empirical Bayesian approach [20]. The basic idea is to marginalize out the unknown Z and E and solve RR max?,? p(Y—Z, E)p(Z—?)p(E—?)dZdE (11) using an EM-like algorithm. Once we have an optimal {?? , ?? }, we then compute the posterior mean of p(Z, E—Y, ?? , ?? ) which is available in closed-form. Finally, a recent class of methods has been derived around the concept of approximate message passing, AMP-RPCA [26], which applies Gaussian priors to the factors A and B and infers posterior estimates by loopy belief propagation [21]. In our experiments (see [23]) we found AMP-RPCA to be quite sensitive to data deviating from these distributions.

3

A New Pseudo-Bayesian Algorithm

As it turns out, it is quite difficult to derive a fully Bayesian model, or some tight variational/empirical approximation, that leads to an efficient algorithm capable of consistently outperforming the original convex PCP, at least in the absence of additional, exploitable prior knowledge. It is here that we adopt 2

Actually many methods attempt to learn this parameter from data, but we avoid this consideration for simplicity. As well, for subtle reasons such learning is sometimes not even identifiable in the strict statistical sense. 3 Note that in [31] this method is motivated from an entirely different variational perspective anchored in convex analysis; however, the cost function that ultimately emerges

is equivalent to what follows with these priors.

3

a pseudo-Bayesian approach, by which we mean that a Bayesian-inspired cost function will be altered using manipulations that, although not consistent with any original Bayesian model, nonetheless produce desirable attributes relevant to blindly solving (1). In some sense however, we view this as a strength, because the final model analysis presented later in Section 4 does not rely on any presumed validity of the underlying prior assumptions, but rather on explicit properties of the objective that emerges, including all assumptions and approximation involved. Basic Model: We begin with the same likelihood function from (6), noting that in the limit as ? ? 0 this will enforce the constraint set from (1). We also adopt the same prior on E given by (7) above and used in [1] and [31], but we need not assume any additional hyperprior on ?. In contrast, for the prior on Z our method diverges, and we define the Gaussian i h ?1 (12) p(Z—?r , ?c ) ? exp ? 21 z ¿ (?r ? I + I ? ?c ) z , where z ,vec[Z] is the column-wise vectorization of Z, ? denotes the Kronecker product, and ?c ?Rn?n and ?r ?Rm?m are positive semi-definite, symmetric matrices.4 Here ?c can be viewed as applying a column-wise covariance factor, and ?r a row-wise one. Note that if ?r =0, then this prior collapses to (10); however, by including ?r we can retain symmetry in our model, or invariance to inference using either Y or Y¿ . Related priors can also be used to improve the performance of affine rank minimization problems [34]. We apply the empirical Bayesian procedure from (11); the resulting convolution of Gaussians integral [2] can be computed in closed-form. After applying ?2 log[?] transformation, this is equivalent to minimizing ? + ?I, (13) L(?r , ?c , ?) = y ¿ ??1 y + log —?y —, where ?y , ?r ? I + I ? ?c + ? y

? , diag[ and ? ? ]. Note that for even reasonably sized problems ?y ? Rnm?nm will be huge, and consequently we will require certain approximations to produce affordable update rules. Fortunately this can be accomplished while simultaneously retaining a principled objective function capable of outperforming existing methods. Pseudo-Bayesian Objective: We first modify (13) to give

P

P L(?r , ?c , ?) = y ¿ ??1 y + j log ?c + 12 ??j + ?2 I + i log ?r + 21 ?i? + ?2 I , (14) y where ??j ,diag[? ?j ] and ? ?j represents the j-th column of ?. Similarly we define ?i? ,diag[? i? ] with ? i? the i-th row of ?. This new cost is nothing more than (13) but with the log — ? — term split in half producing a lower bound by Jensen?s inequality; the Kronecker product can naturally be dissolved under these conditions. Additionally, (14) represents a departure from our original Bayesian model in that there is no longer any direct empirical Bayesian or VB formulation that would lead to (14). Note that although this modification cannot be justified on strictly probabilistic terms, we will see shortly that it nonetheless still represents a viable cost function in the abstract sense, and lends itself to increased computational efficiency. The latter is an immediate effect of the drastically reduced dimensionality of the matrices inside the determinant. Henceforth (14) will represent the cost function that we seek to minimize; relevant properties will be handled in Section 4. We emphasize

that all subsequent analysis is based directly upon (14), and therefore already accounts for the approximation step in advancing from (13). This is unlike other Bayesian model justifications relying on the legitimacy of the original full model, and yet then adopt various approximations that may completely change the problem. Update Rules: Common to many empirical Bayesian and VB approaches, our basic optimization strategy involves iteratively optimizing upper bounds on (14) in the spirit of majorizationminimization [12]. At a high level, our goal will be to apply bounds which separate ?c , ?r , and ? into terms of the general form log —X— + tr[AX?1 ], the reason being that this expression has a simple global minimum over X given by X=A. Therefore the strategy will be to update the bound (parameterized by some matrix A), and then update the parameters of interest X. Using standard conjugate duality relationships and variational bounding techniques [14][Chapter 4], it follows after some linear algebra that 4

Technically the Kronecker sum ?r ?I + I??c must be positive definite for the inverse in (12) to be defined. However, we can accommodate the semi-definite case using the following convention. Without loss of generality assume that ?r ?I + I??c = RR¿ for some matrix R. We then qualify that p(Z—?r , ?c ) = 0 if z ? / span[R], and p(Z—?r , ?c ) ? exp[? 21 z ¿ (R¿ )? R?z ] otherwise.

4

y ¿ ??1 y y

?

1 kY ?

? Z ? Ek2F +

e2 ij i,j ?ij

P

+ z ¿ (?r ? I + I ? ?c )?1 z

(15)

for all Z and E. For fixed values of ?r , ?c , and ? we optimize this quadratic bound to obtain revised estimates for Z and E, noting that exact equality in (15) is possible via the closed-form solution z = (?r ? I + I ? ?c ) ??1 y, y

? ?1 e = ?? y. y

(16)

In large practical problems, (16) may become expensive to compute directly because of the high dimensional inverse involved. However, we may still find the optimum efficiently by an ADMM procedure described in [23]. We can also further bound the righthand side of (15) using Jensen?s inequality as h i ¿ ?1 z ¿ (?r ? I + I ? ?c )?1 z ? tr Z¿ Z??1 . r + ZZ ?c

(17)

Along with (15) this implies that for fixed values of Z and E we can obtain an upper bound which only depends on ?r , ?c , and ? in a decoupled or separable fashion. For the log — ? — terms in (14), we also derive convenient upper bounds using determinant identities and a first-order approximation, the goal being to find a representation that plays well with the previous decoupled bound for optimization purposes. Again using conjugate duality relationships, we can form the bound

log ?c + 12 ??j + ?2 I

? log —?c — + log —??j — + log —W (?c , ??j )— ? log —?c — + log —??j — + tr

h

?1 j (? ?1 )¿ ?c ?

i

c

+ (?c ?1 )¿ ? ?1 ?j +C, ? ?j

(18)

where the inverse ? ?1 ?j is understood to apply element-wise, and W (?c , ??j ) is defined as W (?c , ??j ) ,

1 2?

?

? 2I 2I

2I I

+

??1 c 0

0 ??1 ?j

.

(19)

Additionally, C is a standard constant, which accompanies the first-order approximation to guarantee that the upper bound is tangent to the underlying cost function; however, its exact value is irrelevant for optimization purposes. Finally, the requisite gradients are defined as ?c??1 , ?j

?W (?c ,??j ) ???1 ?j

?j??1 ,

= diag[??j ? 12 ??j (Sjc )?1 ??j ],

c

?W (?c ,??j ) ?1

? ?c

= ?c ? ?c (Sjc )?1 ?c ,

(20) where Sjc , ?c + 21 ??j + ?2 I. Analogous bounds can be derived for the log ?r + 21 ?i? + ?2 I terms in (14). These bounds are principally useful because all ?c , ?r , ??j , and ?i? factors have been decoupled. Consequently, with Z, E, and all the relevant gradients fixed, we can separately combine ?c -, ?r -, and ?-dependent terms from the bounds and then optimize independently. For example, combining terms from (17) and (18) involving ?c for all j, this requires solving hP i j ¿ ?1 ¿ ?1 min m log —?c — + tr (? ) ? + ZZ ? . (21) ?1 c c j ? ?c

c

Analogous cost functions emerge for ?r and ?. All three problems have closed-form optimal solutions given by hP i hP i j¿ 1 1 ¿ i¿ ¿ c + ur , ?c = m ? + ZZ , ? = ? + Z Z , ? = z 2 + u (22) r j i ??1 n ??1 r

c

c , [?c??1 ; . . . ; ?c??1 ], and analogously where the squaring operator is applied element-wise to z , u ?1

?m

1 r . One interesting aspect of (22) is that it forces ?c m for u ZZ¿ and ?r n1 Z¿ Z, thus maintaining a balancing symmetry and preventing one or the other from possibly converging towards zero. This is another desirable consequence of using the bound in (17). To finalize then, the proposed pipeline, which we henceforth refer to as pseudo-Bayesian RPCA (PB-RPCA), involves the steps shown under Algorithm 1 in [23]. These can be implemented in such a way that the complexity is linear in max(n, m) and cubic in min(n, m).

5

4

Analysis of the PB-RPCA Objective

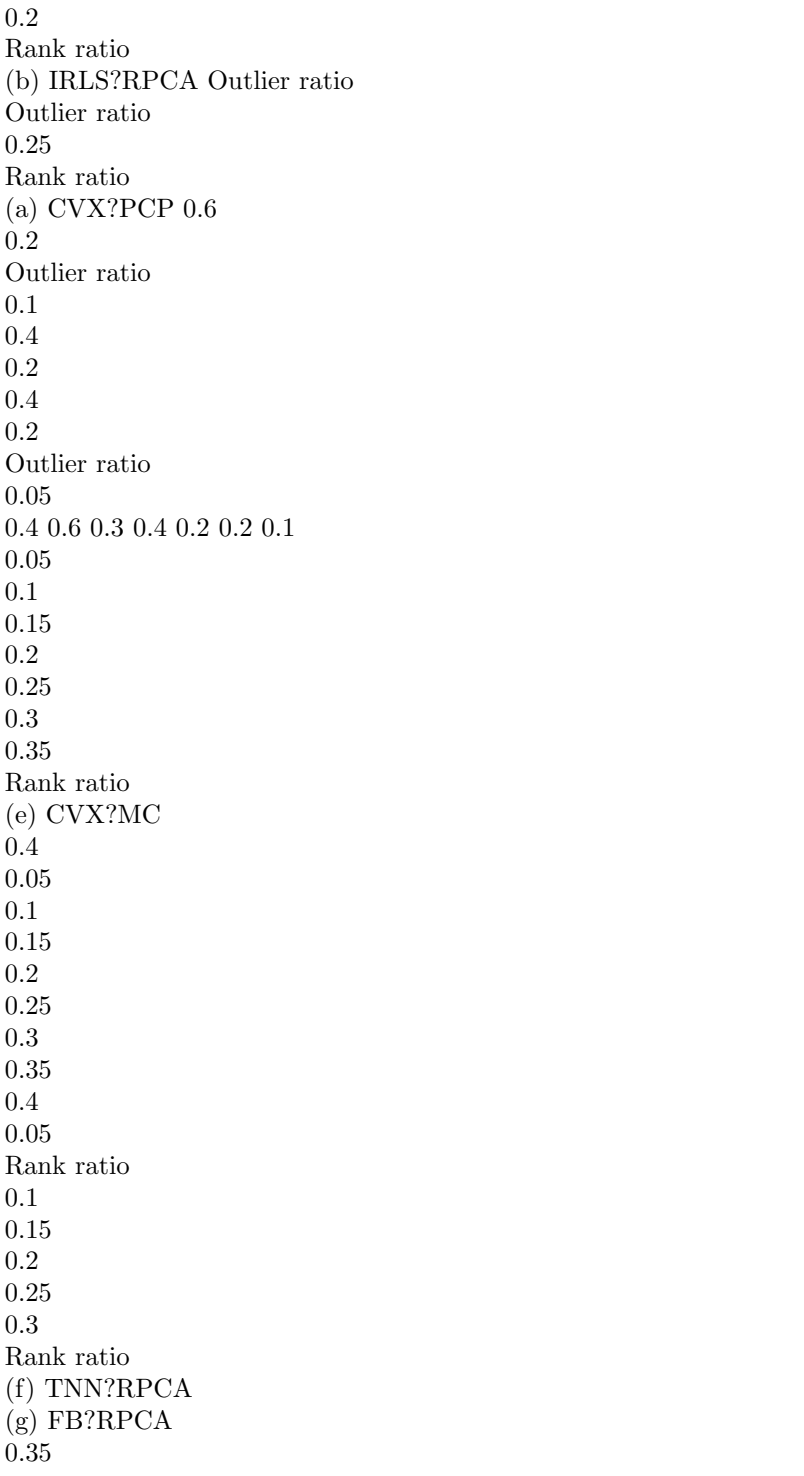
On the surface it may appear that the PB-RPCA objective (14) represents a rather circuitous route to solving (1), with no obvious advantage over the convex PCP relaxation from (3), or any other approach for that matter. However quite surprisingly, we prove in [23] that by simply replacing the log — ? — matrix operators in (14) with tr[?], the resulting function collapses exactly to convex PCP. So what at first appear as distant cousins are actually quite closely related objectives. Of course our work is still in front of us to explain why log — ? —, and therefore the PB-RPCA objective by association, might display any particular advantage. This leads us to considerations of relative concavity, non-separability, and symmetry as described below in turn. Relative Concavity: Although both log — ? — and tr[?] are concave non-decreasing functions of the singular values of symmetric positive definite matrices, and hence favor both sparsity of ? and minimal rank of ?r or ?c , the former is far more strongly concave (in the sense of relative concavity described in [25]). In this respect we may expect that log — ? — is less likely to over-shrink large values [11]. Moreover, applying a concave non-decreasing penalty to elements of ? favors a sparse estimate, ? in (16). which in turn transfers this sparsity directly to E by virtue of the left multiplication by ? Likewise for the singular values of ?c and ?r . Non-Separability: While potentially desirable, the relative concavity distinction described above is certainly not sufficient to motivate why PB-RPCA might represent an effective RPCA approach, especially given the breadth of non-convex alternatives already in the literature. However, a much stronger argument can be made by exposing a fundamental limitation of all RPCA methods (convex or otherwise) that rely on minimization of generic penalties in the separable or additive form of (4). For this purpose, let ? denote a set of indices that correspond with zero-valued elements in E, such that E? = 0 while all other elements of E are arbitrary nonzeros (it can equally be viewed as the complement of the support of E). In the case of MC, ? would also represent the set of observed matrix elements. We then have the following: Proposition 1. To guarantee that (4) has the same global optimum as (1) for all Y where a unique solution exists, it follows that f1 and f2 must be non-convex and no feasible descent direction can ever remove an index from or decrease the cardinality of ?. In [31] it has been shown that, under similar conditions, the gradient in a feasible direction at any zero-valued element of E must be infinite to guarantee a matching global optimum, from which this result naturally follows. The ramifications of this

proposition are profound if we ever wish to produce a version of RPCA that can mimic the desirable behavior of much simpler MC problems with known support, or at least radically improve upon PCP with unknown outlier support. In words, Proposition 1 implies that under the stated global-optimality preserving conditions, if any element of E converges to zero during optimization with an arbitrary descent algorithm, it will remain anchored at zero until the end. Consequently, if the algorithm prematurely errs in setting the wrong element to zero, meaning the wrong support pattern has been inferred at any time during an optimization trajectory, it is impossible to ever recover, a problem naturally side-stepped by MC where the support is effectively known. Therefore, the adoption of separable penalty functions can be quite constraining and they are unlikely to produce sufficiently reliable support recovery. But how does this relate to PB-RPCA? Our algorithm maintains a decidedly non-separable penalty function on ?c , ?r , and ?, which directly transfers to an implicit, non-separable regularizer over Z and E when viewed through the dual-space framework from [32].5 By this we mean a penalty f (Z, E)6=f1 (Z)+f2 (E) for any functions f1 and f2 , and with Z fixed, we have P f (Z, E)6= i,j fij (eij ) for any set of functions {fij }. We now examine the consequences. Let ? now denote a set of indices that correspond with zerovalued elements in ?, which translates into an equivalent support set for Z via (16). This then leads to quantifiable benefits: Proposition 2. The following properties hold w.r.t. the PB-RPCA objective (assuming n = m for simplicity): ? Assume that a unique global solution to (1) exists such that either rank[Z]+maxj ke?j k0 ¡n or rank[Z]+maxi kei? k0 ¡n. Additionally, let {??c , ??r , ?? } denote a globally minimizing solution to (14) and {Z? ,E? } the corresponding values of Z and E computed using (16). Then in the limit ??0, Z? and E? globally minimize (1). 5

Even though this penalty function is not available in closed-form, non-separability is nonetheless enforced via the linkage between ?c , ?r , and ? in the log — ? — operator.

6
1
0.6
0.6
0.6
0.5
0.6
0.2
Outlier ratio
0.4
0.4
0.2
0.4
0.2
Outlier ratio
0.2
Outlier ratio

Outlier ratio
Outlier ratio
0.8
0.4
0.4 0.6 0.3 0.4 0.2 0.2 0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.05
0.1
0.15
Rank ratio
0.3
0.35
0.4
0.05
0.1
0.15
[Known outlier location]
0.2
0.3
0.35
0.4
0.05
0.1 0.05
0.4
0.2
0.15 0.150.2 0.2 0.25 0.1 0.25
0.3 0.3
0.35 0.40.4 0.35
0
Rank Rank ratioratio
(d) PB?RPCA w/o sym. 1
0.6
0.5
0.6
0.8
Outlier ratio
0.4
0.25
(c) VB?RPCA
[Known rank]
0.6


0.2
Rank ratio
(b) IRLS?RPCA Outlier ratio
Outlier ratio
0.25
Rank ratio
(a) CVX?PCP 0.6
0.2
Outlier ratio
0.1
0.4
0.2
0.4
0.2
Outlier ratio
0.05
0.4 0.6 0.3 0.4 0.2 0.2 0.1
0.05
0.1
0.15
0.2
0.25
0.3
0.35
Rank ratio
(e) CVX?MC
0.4
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.05
Rank ratio
0.1
0.15
0.2
0.25
0.3
Rank ratio
(f) TNN?RPCA
(g) FB?RPCA
0.35

0.4
0.05
0.1 0.05
0.15 0.150.2 0.2 0.25 0.1 0.25
0.3 0.3
0.35 0.40.4 0.35
0
Rank Rank ratioratio
(h) PB?RPCA (Proposed)

Figure 1: Phase transition over outlier (y-axis) and rank (x-axis) ratio variations. Here CVX-MC and TNN-RPCA maintain advantages of exactly known outlier support pattern and true rank respectively. ? Assume that Y has no entries identically equal to zero.6 Then for any arbitrary ?, there will always exist a range of ?c and ?r values such that for any ? consistent with ? we are not at a locally minimizing solution to (14), meaning there exists a feasible descent direction whereby elements of ? can escape from zero. A couple important comments are worth stating regarding this result. First, the rank and row/columnsparsity requirements are extremely mild. In fact, any minimum of (1) will be such that rank[Z] + maxj ke?j k0 ? n and rank[Z] + maxi kei? k0 ? m, regardless of Y. Secondly, unlike any separable penalty function (4) that retains the correct global optimal as (1), Proposition 2 implies that (14) need not be locally minimized by every possible support pattern for outlier locations. Consequently, premature convergence to suboptimal supports need not disrupt trajectories towards the global solution to the extent that (4) may be obstructed. Moreover, beyond algorithms that explicitly adopt separable penalties (the vast majority), some existing Bayesian approaches may implicitly default to (4). For example, as shown in [23], the mean-field factorizations adopted by VB-RPCA actually allow the underlying free energy objective to be expressible as (4) for some f1 and f2 . Symmetry: Without the introduction of symmetry via our pseudo-Bayesian proposal (meaning either ?c or ?r is forced to zero), then PB-RPCA collapses to something like EB-RPCA, which depends heavily on whether Y or Y¿ is provided as input and penalizes column- and row-spaces asymmetrically. In this regime it can be shown that the analogous requirement to replicate Proposition 2 becomes more stringent, namely we must assume the asymmetric condition rank[Z] + maxj ke?j k0 ¡ n. Thus the symmetric cost of PB-RPCA of allows us to relax this column-wise restriction provided a row-wise alternative holds (and vice versa), allowing the PB-RPCA objective (14) to match the global optimum of our original problem from (1) under broader conditions. In closing this section, we reiterate that all of our analysis and conclusions are based on (14), after the stated approximations. Therefore we need not rely on the plausibility of the original Bayesian starting point from Section 3 nor the tightness of subsequent approximations for justification; rather (14) can be viewed as a principled stand-alone objective for RPCA regardless of its origins. Moreover, it represents the first approach satisfying the relative concavity, non-separability, and symmetry properties described above, which can loosely be viewed as necessary, but not sufficient design criteria for an optimal

RPCA objective.

5

Experiments

To examine significant factors that influence the ability to solve (1), we first evaluate the relative performance of PB-RPCA estimating random simulated subspaces from corrupted measurements, the standard benchmark. Later we present subspace clustering results for motion segmentation as a practical application. Additional experiments and a photometric stereo example are provided in [23]. Phase Transition Graphs: We compare our method against existing RPCA methods: PCP [16], TNN [24], IRLS [18], VB [1], and FB [9]. We also include results using PB-RPCA but with symmetry removed (which then defaults to something like EB-RPCA), allowing us to isolate the importance of this factor, called ?PB-RPCA w/o sym.?. For competing algorithms, we set parameters based on the values suggested by original authors with the exception of IRLS. Detailed settings and parameters can be found in [23]. 6

This assumption can be relaxed with some additional effort but we avoid such considerations here for clarity of presentation.

7

?

Success Rate

1 0.8 0.6 0.4

PB-RPCA (easy case) PB-RPCA (hard case) PCP (easy case) PCP (hard case)

0.2 0 0

0.2

0.4 0.6 Outlier Ratio

SSC Robust SSC PCP+SSC PB+SSC (Ours) Without sub-sampling (large number of measurements) 19.0 / 14.9 5.3 / 0.3 3.0 / 0.0 2.4 / 0.0 28.2 / 28.3 6.4 / 0.4 3.0 / 0.0 2.4 / 0.0 33.2 / 34.7 7.2 / 0.5 3.6 / 0.2 2.8 / 0.0 36.5 / 39.0 8.5 / 0.6 4.7 / 0.2 3.1 / 0.0 With sub-sampling (small number of measurements) 0.1 19.5 / 17.2 4.0 / 0.0 2.9 / 0.0 2.8 / 0.0 0.2 33.0 / 33.3 5.3 / 0.0 3.7 / 0.0 3.6 / 0.0 0.3 39.3 / 41.1 5.7 / 1.7 5.0 / 0.7 3.9 / 0.0 42.2 / 43.5 6.4 / 2.1 9.8 / 5.1 3.7 / 0.0 0.4 *Values are percentage with (mean / median). 0.1 0.2 0.3 0.4

0.8

Figure 2: Hard case comparison.

1

Figure 3: Motion segmentation errors on Hopkins155.

We construct phase transition plots as in [4, 9] that evaluate the recovery success of every pairing of outlier ratio and rank using data Y=ZGT +EGT , where Y?Rm?n and m=n=200. The ground truth outlier matrix EGT is generated by selecting non-zero entries uniformly with probability ??[0,1], and its magnitudes are sampled iid from the uniform distribution U [?20, 20]. We generate the ground truth low-rank matrix by ZGT =AB¿ , where A?Rn?r and B?Rm?r are drawn from iid N (0,1). Figure 1 shows comparisons among competing methods, as well as the convex nuclear norm based matrix completion (CVX-MC) [5], the latter representing a far easier estimation task given that

missing entry locations (analogous to corruptions) occur in known locations. The color of each cell encodes the percentage of success trials (out of 10 total) whereby the normalized root-mean-squared ? GT kF error (NRMSE, kZ?Z ) recovering ZGT is less than 0.001 to classify success following [4, 9]. kZGT kF Notably PB-RPCA displays a much broader recoverability region. This improvement is even maintained over TNN-RPCA and MC which require prior knowledge such as the true rank and exact outlier locations respectively. These forms of prior knowledge offer a substantial advantage, although in practical situations are usually unavailable. PB-RPCA also outperforms PB-RPCA w/o sym. (its closest relative) by a wide margin, suggesting that the symmetry plays an important role. The poor performance of FB-RPCA is explained in [23]. Hard Case Comparison: Recovery of Gaussian iid low-rank components (the typical benchmark recovery problem in the literature) is somewhat ideal for existing algorithms like PCP because the singular vectors of ZGT will not resemble unit vectors that could be mistaken for sparse components. However, a simple test reveals just how brittle PCP is to deviations from the theoretically optimal regime. We generate a rank one ZGT = ?a3 (b3 )¿ , where the cube operation is applied element-wise, a and b are vectors drawn iid from a unit sphere, and ? scales ZGT to unit variance. EGT has nonzero elements drawn iid from U [?1, 1]. Figure 2 shows the recovery results as the outlier ratio is increased. The hard case refers to the data just described, while the easy case follows the model used to make the phase transition plots. While PB-RPCA is quite stable, PCP completely fails for the hard data. Outlier Removal for Motion Segmentation: Under an affine camera model, the stacked matrix consisting of feature point trajectories of k rigidly moving objects forms a union of k affine subspaces of at most rank 4k [29]. But in practice, mismatches often occur due to occlusions or tracking algorithm limitations, and these introduce significant outliers into the feature motions such that the corresponding trajectory matrix may be at or near full rank. We adopt an experimental paradigm from [17] designed to test motion segmentation estimation in the presence of outliers. To mimic mismatches while retaining access to ground-truth, we randomly corrupt the entries of the trajectory matrix formed from Hopkins155 data [28]. Specifically, following [17] we add noise drawn from N (0, 0.1?) to randomly sampled points with outlier ratio ??[0, 1], where ? is the maximum absolute value of the data. We may then attempt to recover a clean version from the corrupted measurements using RPCA as a preprocessing step; motion segmentation can then be applied using standard subspace clustering [29]. We use SSC and robust SSC algorithms [10] as baselines, and compare with RPCA preprocessing computed via PCP (as suggested in [10]) and PB-RPCA followed by SSC. Additionally, we sub-sampled the trajectory matrix to increase problem difficulty by fewer samples. Segmentation accuracy is reported in Fig. 3, where we observe that PB shows the best performance across different outlier ratios, and the performance gap widens when the measurements are scarce.

6

Conclusion

Since the introduction of convex RPCA algorithms, there has not been a

significant algorithmic break-through in terms of dramatically enhancing the regime where success is possible, at least in the absence of any prior information (beyond the generic low-rank and sparsity assumptions). The likely explanation is that essentially all of these approaches solve either a problem in the form of (4), an asymmetric problem in the form of (11), or else require strong priori knowledge. We provide a novel integration of three important design criteria, concavity, non-separability, and symmetry, that leads to state-of-the-art results by a wide margin without tuning parameters or prior knowledge. 8

# 2 References

[1] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos. Sparse Bayesian methods for low-rank matrix estimation. IEEE Trans. Signal Process., 2012. [2] C. M. Bishop. Pattern recognition and machine learning. Springer New York, 2006. [3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning R in Machine Learning, 2011. via the alternating direction method of multipliers. Foundations and Trends [4] E. J. Cand?s, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? J. of the ACM, 2011. [5] E. J. Cand?s and B. Recht. Exact matrix completion via convex optimization. Foundations of Computational mathematics, 2009. [6] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. SIAM J. on Optim., 2011. [7] R. Chartrand. Nonconvex splitting for regularized low-rank+ sparse decomposition. IEEE Trans. Signal Process., 2012. [8] Y.-L. Chen and C.-T. Hsu. A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In IEEE Int. Conf. Comput. Vis., 2013. [9] X. Ding, L. He, and L. Carin. Bayesian robust principal component analysis. IEEE Trans. Image Process., 2011. [10] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. IEEE Trans. Pattern Anal. and Mach. Intell., 2013. [11] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc., 2001. [12] D. R. Hunter and K. Lange. A tutorial on MM algorithms. The American Statistician, 2004. [13] H. Ji, C. Liu, Z. Shen, and Y. Xu. Robust video denoising using low rank matrix completion. In IEEE Conf. Comput. Vis. and Pattern Recognit., 2010. [14] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. Mach. Learn., 1999. [15] B. Lakshminarayanan, G. Bouchard, and C. Archambeau. Robust Bayesian matrix factorisation. In AISTATS, 2011. [16] Z. Lin, M. Chen, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv:1009.5055, 2010. [17] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In IEEE Int. Conf. Comput. Vis., 2011. [18] C. Lu, Z. Lin, and S. Yan. Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. IEEE Trans. Image Process., 2015. [19] K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. J. Mach. Learn.

Res., 2012. [20] K. P. Murphy. Machine Learning: a Probabilistic Perspective. MIT Press, 2012. [21] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In UAI, 1999. [22] T.-H. Oh, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon. Robust high dynamic range imaging by rank minimization. IEEE Trans. Pattern Anal. and Mach. Intell., 2015. [23] T.-H. Oh, Y. Matsushita, I. S. Kweon, and D. Wipf. Pseudo-Bayesian robust PCA: Algorithms and analyses. arXiv preprint arXiv:1512.02188, 2015. [24] T.-H. Oh, Y.-W. Tai, J.-C. Bazin, H. Kim, and I. S. Kweon. Partial sum minimization of singular values in Robust PCA: Algorithm and applications. IEEE Trans. Pattern Anal. and Mach. Intell., 2016. [25] J. A. Palmer. Relative convexity. ECE Dept., UCSD, Tech. Rep, 2003. [26] J. T. Parker, P. Schniter, and V. Cevher. Bilinear generalized approximate message passing. arXiv:1310.2632, 2013. [27] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. IEEE Trans. Pattern Anal. and Mach. Intell., 2012. [28] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In IEEE Conf. Comput. Vis. and Pattern Recognit., 2007. [29] R. Vidal. Subspace clustering. IEEE Signal Process. Mag., 2011. [30] N. Wang and D.-Y. Yeung. Bayesian robust matrix factorization for image and video processing. In IEEE Int. Conf. Comput. Vis., 2013. [31] D. Wipf. Non-convex rank minimization via an empirical Bayesian approach. In UAI, 2012. [32] D. Wipf, B. D. Rao, and S. Nagarajan. Latent variable Bayesian models for promoting sparsity. IEEE Trans. on Information Theory, 2011. [33] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma. Robust photometric stereo via low-rank matrix completion and recovery. In Asian Conf. Comput. Vis., 2010. [34] B. Xin and D. Wipf. Pushing the limits of affine rank minimization by adapting probabilistic PCA. In Int. Conf. Mach. Learn., 2015.

9