Bilinear Parameterization For Differentiable Rank-Regularization

Marcus Valtonen Örnhag[1] Carl Olsson[1,2] Anders Heyden[1]

[1]Centre for Mathematical Sciences
Lund University

[2]Department of Electrical Engineering
Chalmers University of Technology

{marcus.valtonen_ornhag, carl.olsson, anders.heyden}@math.lth.se

Abstract

Low rank approximation is a commonly occurring problem in many computer vision and machine learning applications. There are two common ways of optimizing the resulting models. Either the set of matrices with a given rank can be explicitly parametrized using a bilinear factorization, or low rank can be implicitly enforced using regularization terms penalizing non-zero singular values. While the former approach results in differentiable problems that can be efficiently optimized using local quadratic approximation, the latter is typically not differentiable (sometimes even discontinuous) and requires first order subgradient or splitting methods. It is well known that gradient based methods exhibit slow convergence for ill-conditioned problems.

In this paper we show how many non-differentiable regularization methods can be reformulated into smooth objectives using bilinear parameterization. This allows us to use standard second order methods, such as Levenberg–Marquardt (LM) and Variable Projection (VarPro), to achieve accurate solutions for ill-conditioned cases. We show on several real and synthetic experiments that our second order formulation converges to substantially more accurate solutions than competing state-of-the-art methods.[1]

1. Introduction

Low rank models have been applied to numerous vision applications ranging from high level shape and deformation to pixel appearance models [48, 6, 52, 22, 2, 21, 50, 11]. When the sought rank is known, a commonly occurring formulation is the least squares minimization

$$\min_{\text{rank}(X)\leq r} \|\mathcal{A}X - b\|^2, \tag{1}$$

where $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ is a linear operator, and $\|\cdot\|$ is the standard Euclidean vector norm. In general, this is a difficult non-convex problem and some versions are even known to be NP-hard [26]. In structure from motion, a popular approach [7] is to optimize over a bilinear factorization $X = BC^T$, where B is $m \times r$ and C is $n \times r$, and solve

$$\min_{B,\,C} \|\mathcal{A}BC^T - b\|^2. \tag{2}$$

Since the rank is bounded by the number of columns in B and C this approach explicitly parametrizes the set of matrices of rank r. While bilinear approaches often perform well [29, 16] they can have local minima [7]. Recent works [29, 30, 31, 33] have, however, shown that properly implemented, LM and VarPro approaches are remarkably robust to local minima, achieve quadratic convergence and give impressive reconstruction results. Recently [24, 3, 23] was able to give conditions which guarantee that there are no "spurious" local minimizers (meaning that all local minimizers are close to or identical to the global solution). They use the notion of restricted isometry property (RIP) [45] which assumes that the operator $\mathcal{A}$ fulfills

$$(1-\delta_r)\|X\|_F^2 \leq \|\mathcal{A}X\|^2 \leq (1+\delta_r)\|X\|_F^2, \tag{3}$$

with $0 \leq \delta_r < 1$, if $\text{rank}(X) \leq r$. If the isometry constant δ_r is sufficiently small [24, 23, 3] prove that every local minimizer is optimal (or near optimal). Similarly, for the matrix completion problem [23] showed that there are no spurious local minima under uniformly distributed missing data. While the above theoretical assumptions generally do not hold for computer vision problems such as structure from motion, these results still give some intuition as to why bilinear parameterization often works well.

An alternative approach is to optimize directly over the entries of X and penalize high rank using regularization terms. Applying a robust function f to the singular values $\sigma_i(X) = 1, \dots, N = \min(m, n)$ results in a low-rank inducing objective

$$\min_X \mathcal{R}(X) + \|\mathcal{A}X - b\|^2, \tag{4}$$

where $\mathcal{R}(X) = \sum_{i=1}^N f(\sigma_i(X))$. Besides controlling the rank of the solution the generality of the function f offers

[1]This work was supported by the Swedish Research Council (grants no. 2015-05639 and 2018-05375) and the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots).

increased modeling capability compared to (1) and can for example be used to add priors on the size of the non-zero singular values.

The most popular regularization approach is undoubtedly the nuclear norm, $f(\sigma_i(X)) = \sigma_i(X)$, due to its convexity [18, 45, 44, 9, 10]. Under the RIP assumption exact or approximate recovery with the nuclear norm can then be guaranteed [45, 10]. On the other hand, since it penalizes large singular values, it suffers from a shrinking bias [8, 11, 35]. Ideally f should penalize small singular values (assumed to stem from measurement noise) harder than the large ones. Therefore non-increasing derivatives on $[0, \infty)$, or concavity, has been shown to give stronger relaxations [43, 37, 32, 39, 12, 47, 27]. These non-convex formulations usually only come with local convergence guarantees. Two exceptions are [35, 40] which gave optimality guarantees for (4) with $f = f_\mu$ as in (8).

The regularization term is generally not differentiable as a function of X. Thus, optimization methods based on local quadratic approximation become infeasible. Figure 1 gives a simple illustration of a 1-dimensional example of how non-differentiability occurs at the origin. In addition it is well known that the singular values become non-differentiable functions of the matrix elements when they are non distinct. To circumvent these issues subgradient and splitting methods are often employed [12, 47, 27, 38, 35]. It is well known from basic optimization theory (*e.g.* [5]) that gradient based methods exhibit slow convergence for ill-conditioned problems. It has also been observed (*e.g.* [4]) that splitting methods rapidly reduce the objective value the first couple of iterations, while convergence to the exact solution can be slow. In this paper we show that there are computer vision problems where these approaches make very little improvements at all, returning a solution that is far from optimal. In contrast, bilinear formulations with either LM or VarPro can be made to yield accurate results in few iterations [29].

An alternative approach that unifies bilinear parameterization with regularization approaches is based on the observation [45] that the nuclear norm $\|X\|_*$ of a matrix X can be expressed as $\|X\|_* = \min_{BC^T=X} \frac{\|B\|_F^2+\|C\|_F^2}{2}$. Thus when $f(\sigma_i(X)) = \mu\sigma_i(X)$, where μ is a scalar controlling the strength of the regularization, optimization of (4) can be formulated as

$$\min_{B,C} \mu\frac{\|B\|_F^2 + \|C\|_F^2}{2} + \|\mathcal{A}BC^T - b\|^2. \tag{5}$$

Optimizing directly over the factors has the advantages that the number of variables is much smaller and one may add constraints if a particular factorization is sought. Surprisingly, while (5) is non-convex, using the convexity of the underlying regularization problem (4) it can be shown that any local minimizer B,C with $\mathrm{rank}(BC^T) < k$, where k is the number of columns in B and C, is globally optimal [1, 28]. Additionally, the objective function is two times differentiable and second order methods can be employed.

In this paper we develop new regularizing terms that, similar to (5), work on the bilinear factors. However, in contrast to previous approaches we investigate formulations that exhibit less shrinking bias and go beyond convex penalties. Specifically, we prove that $\mathcal{R}(X) = \min_{X=BC^T} \tilde{\mathcal{R}}(B, C)$, where

$$\tilde{\mathcal{R}}(B, C) = \sum_{i=1}^{k} f\left(\frac{\|B_i\|^2 + \|C_i\|^2}{2}\right), \tag{6}$$

k is the number of columns, and B_i and C_i are the i:th columns of B and C, respectively. The result holds for a general class of concave penalty functions f, a few of which are illustrated in Figure 1. In view of the above result, we propose to minimize

$$\tilde{\mathcal{R}}(B, C) + \|\mathcal{A}BC^T - b\|^2. \tag{7}$$

Rather than resorting to splitting or subgradient methods we present an algorithm that uses a quadratic approximation of the objective. Under the assumption that f is differentiable, we show that our quadratic approximation reduces to a weighted version of (5) to which we can apply VarPro. We show on several computer vision problems that our approach outperforms state-of-the-art methods such as [46, 12, 47, 27, 4].

While our problem is non-convex (both in the X parameterization (4) and in the B, C parameterization (7)) we show that in some cases it is still possible to give global optimality guarantees. Building on the results of [40] we characterize the local minima of the new formulation with the choice

$$f(x) = f_\mu(x) := \mu - \max(\sqrt{\mu} - x, 0)^2. \tag{8}$$

Specifically, for this choice, we give conditions that ensure that when a RIP constraint [45] holds a local minimizer of (7) is a global solution of both

$$\min_{\mathrm{rank}(X)\leq r} \mathcal{R}(X) + \|\mathcal{A}X - b\|^2, \tag{9}$$

where $\mathcal{R}(X) = \sum_i f_\mu(\sigma_i(X))$, and

$$\min_{\mathrm{rank}(X)\leq r} \mu\mathrm{rank}(X) + \|\mathcal{A}X - b\|^2. \tag{10}$$

In summary our main contributions are:

- A new stronger non-convex regularization term for bilinear parameterizations with less/no shrinking bias.
- A new iteratively reweighed VarPro algorithm optimizing accurate quadratic approximations.
- Theoretical conditions that guarantee optimal recovery under the RIP constraint.
- An experimental evaluation that shows that our methods outperforms state-of-the-art methods on several real computer vision problems.

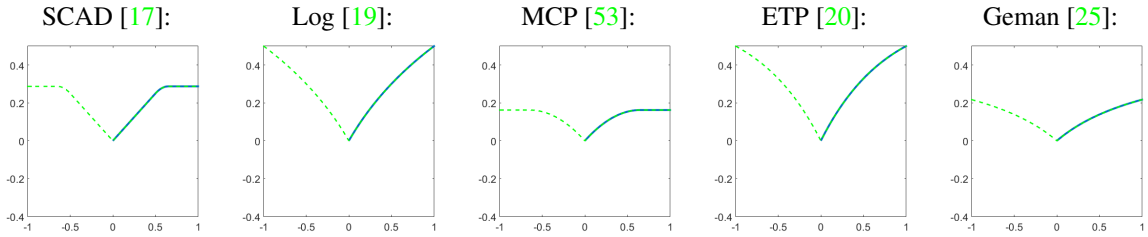

Figure 1. A few commonly occurring robust penalties of the form $f(\sigma)$, with $\sigma \in [0, \infty)$ and f differentiable everywhere (blue graph). The green dashed graph shows how non-differentiability occurs at the origin when applying the penalty to a 1×1 matrix $x \in \mathbb{R}$. In this case $\sigma(x) = |x|$ and therefore $f(\sigma(x)) = f(|x|)$. Note also that (8) is a special case of MCP.

1.1. Related Work

Our work is very much inspired by a recent series of papers by Hong *et al*. [29, 30, 31, 33] which show that bilinear formulations can be made remarkably robust to local minima, and achieve impressive reconstruction results for uncalibrated structure from motion problems, using the so called VarPro method. Our work represents an attempt to unify this line of work with regularization based alternatives, leveraging the benefits of them both.

An approach that is closely related to ours is that of [8] which uses (5) to unify the use of a regularized objective and factorization. They show that if the obtained solution has lower rank than its number of columns it is globally optimal. In practice [8] observes that the shrinking bias of the nuclear norm makes it too weak to enforce a low rank when the data is noisy. Therefore, a "continuation" approach where the size of the factorization is gradually reduced is proposed. While this yields solutions with lower rank, the optimality guarantees no longer apply.

Bach *et al*. [1] showed that

$$\|X\|_{s,t} := \min_{X=BC^T} \sum_{i=1}^{k} \frac{\|B_i\|_s^2 + \|C_i\|_t^2}{2}, \qquad (11)$$

is convex for any choice of vector norms $\|\cdot\|_s$ and $\|\cdot\|_t$. In [28] it was shown that a more general class of 2-homogeneous factor penalties result in a convex regularization similar to (11). The property that a local minimizer B, C with $\operatorname{rank}(BC^T) < k$ is global, is also extended to this case. Still, because of convexity, it is clear that these formulations will suffer from a similar shrinking bias as the nuclear norm. Shang *et al*. [46] showed that penalization with the Schatten semi-norms $\|X\|_q = \sqrt[q]{\sum_{i=1}^{N} \sigma_i(X)^q}$, for $q = 1/2$ and $2/3$, can be achieved using a convex penalty on the factors B and C. A generalization to general values of q is given in [51]. While this reduces shrinking bias to some extent, it results in a non-differentiable and non-convex formulation that is optimized with ADMM.

In [34] a bilinear framework on the class of semi-definite matrices was proposed for a general class of low-rank inducing penalties.

Valtonen Örnhag *et al*. [42] considered a framework similar to the one we propose; however, the rank is assumed to be known a priori. Furthermore, they do not show equivalence between the proposed bilinear regularizer and the corresponding original fixed rank regularizer from [35].

It is important to note that many of the above methods that are considered state-of-the-art have been developed for low level vision tasks such as image denoising, inpainting, alignment and background subtraction. The ground truth for these models are often of higher rank than models in *e.g*. structure from motion, making it possible to obtain good results with weaker regularization. Additionally, as we will see in the experiments, more difficult data terms prevent rapid convergence of the splitting methods they often employ.

2. Non-Convex Penalties and Shrinking Bias

In this section we will show how to formulate regularization terms of the type

$$\mathcal{R}(X) = \sum_{i=1}^{N} f(\sigma_i(X)), \qquad (12)$$

by penalizing the factors of the factorization $X = BC^T$. We assume that B and C have k columns, making $\sigma_i(X) = 0$ if $i > k$ and $\operatorname{rank}(X) \leq k$. Note, however, that we are aiming to achieve a lower rank using the regularization term. In many applications, the sought rank is unknown and should be determined by the regularization. We therefore set k large enough not to exclude the optimal solution. As we shall see in Section 3, this ability to over-parameterize can be used to ensure optimality.

Theorem 1. *If f is concave, non-decreasing on $[0, \infty)$ and $f(0) = 0$ then*

$$\mathcal{R}(X) = \min_{BC^T=X} \sum_{i=1}^{k} f(\|B_i\|\|C_i\|), \qquad (13)$$

where B_i and C_i, $i = 1, ..., k$ are the columns of B and C respectively.

Proof. The result is a consequence of the fact that $\mathcal{R}$ will fulfill a triangle inequality $\mathcal{R}(X + Y) \leq \mathcal{R}(X) + \mathcal{R}(Y)$

under the assumptions on f. This is clear from Theorem 4.4 in [49] which shows that

$$\sum_{i=1}^{N} f(\sigma_i(X+Y)) \leq \sum_{i=1}^{N}(f(\sigma_i(X)) + \sum_{i=1}^{N} f(\sigma_i(Y))). \tag{14}$$

Applying this to $X = BC^T = \sum_{i=1}^{k} B_i C_i^T$ we see that

$$\mathcal{R}(X) = \mathcal{R}(\sum_{i=1}^{k} B_i C_i^T) \leq \sum_{i=1}^{k} \mathcal{R}(B_i C_i^T). \tag{15}$$

Since $\operatorname{rank}(B_i C_i^T) = 1$ we also have

$$\mathcal{R}(B_i C_i^T) = f(\sigma_1(B_i C_i^T)) = f(\|B_i C_i^T\|_F). \tag{16}$$

Lastly, since $\|B_i C_i^T\|_F = \|B_i\|\|C_i\|$ we get

$$\mathcal{R}(X) \leq \sum_{i=1}^{k} f(\|B_i\|\|C_i\|). \tag{17}$$

To see that equality can be achieved, let $B_i = \sqrt{\sigma_i(X)}U_i$ and $C_i = \sqrt{\sigma_i(X)}V_i$, where $X = \sum_{i=1}^{k} \sigma_i(X) U_i V_i^T$ is the SVD of X. Then, $BC^T = X$ and $f(\|B_i\|\|C_i\|) = f(\sigma_i(X))$. □

While the above result allows optimization over the factors B and C we note that it yields an objective that is non-differentiable at $\|B_i\|\|C_i\| = 0$. Next we reformulate the objective to achieve a differentiable problem formulation.

Corollary 1. *Under the assumptions of Theorem 1, it follows that* $\mathcal{R}(X) = \min_{X=BC^T} \tilde{\mathcal{R}}(B, C)$*, where*

$$\tilde{\mathcal{R}}(B, C) = \sum_{i=1}^{k} f\left(\frac{\|B_i\|^2 + \|C_i\|^2}{2}\right). \tag{18}$$

If f is differentiable then $\tilde{R}(B, C)$ is also differentiable.

Proof. By the rule of arithmetic and geometric means

$$\|B_i\|\|C_i\| \leq \frac{1}{2}(\|B_i\|^2 + \|C_i\|^2), \tag{19}$$

with equality if $\|B_i\| = \|C_i\|$ which is achieved when $B_i = \sqrt{\sigma_i(X)}U_i$ and $C_i = \sqrt{\sigma_i(X)}V_i$. Since f is assumed to be non-decreasing, it follows from (13), that $\mathcal{R}(X) = \min_{X=BC^T} \tilde{\mathcal{R}}(B, C)$. The differentiability of $\tilde{\mathcal{R}}(B, C)$ is now trivially checked using the chain rule. □

We are particularly interested in the case (8) since, with this choice, it is known that the global minimizer of (4) is the same as that of $\mu\operatorname{rank}(X) + \|\mathcal{A}X - b\|^2$ if $\|\mathcal{A}\| < 1$, see [13] for a proof. Note that f_μ is a special case of the MCP class [53]. With this choice $\tilde{\mathcal{R}}(B, C)$ is differentiable and the second derivatives are also defined almost everywhere except in the transition $\frac{\|B_i\|^2 + \|C_i\|^2}{2} = \sqrt{\mu}$ where the function switches from quadratic to constant.

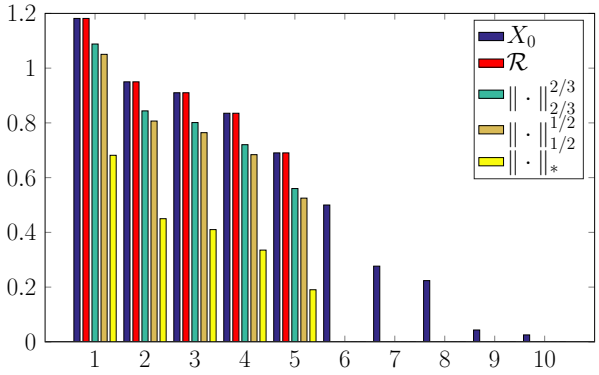

Figure 2. Singular values obtained when minimizing $\|X - X_0\|_F^2$ with the four regularizers $\mathcal{R}(X)$ with $f = f_\mu$, $\|X\|_{1/2}^{1/2}$, $\|X\|_{2/3}^{2/3}$ and $\|X\|_*$. Large singular values are left unchanged by $\mathcal{R}$.

We conclude this section by comparing the shrinking bias of our approach and three others that can also be optimized over the factorization. Theorem 1 makes it possible to compute the global optimizer of $\tilde{\mathcal{R}}(B, C) + \|BC^T - X_0\|_F^2$ since the equivalent problem $\mathcal{R}(X) + \|X - X_0\|_F^2$ has closed form solution in the X-parameterization. It is shown in [35] that with $f = f_\mu$ the solution is obtained by thresholding the singular values at $\sqrt{\mu}$. Similarly, closed form solutions are also available when regularizing $\|X - X_0\|_F^2$ with $\|\cdot\|_{1/2}$, $\|\cdot\|_{2/3}$ and $\|\cdot\|_*$ [46]. In Figure 2 we show the singular values obtained when regularizing $\|X - X_0\|_F^2$ with these four options, and for comparison the singular values of X_0. For all methods we have selected regularization weights as small as possible so that the five smallest singular values are completely suppressed, which minimizes the bias. While all choices, except $\mathcal{R}$, subtract a part from the singular values that should be retained, the Schatten norms reduce the bias significantly compared to the nuclear norm. For the Schatten norms the bias is larger for singular values that are close to the threshold since the derivative of σ^q, $0 < q < 1$, decreases with increasing σ. For problem instances where there is a clear separation in size between singular values that should be retained and those that should be suppressed, it is likely that this can be done with negligible bias. Since $f'_\mu(\sigma) = 0$ when $\sigma \geq \sqrt{\mu}$ this method does not affect the first five singular values.

3. Overparameterization and Optimality

The results of the previous section show that a global optimizer (B, C) of (7) gives a solution BC^T which is globally optimal in (4). On the other hand, optimizing (7) over B and C introduces additional stationary points, due to the non-linear parameterization, that are not present in (4). One such point is $(B, C) = (0, 0)$ where the gradients of $\|\mathcal{A}BC^T - b\|^2$ with respect to B and C vanish (in contrast

Table 1. Distance to ground truth (normalized) mean valued over 20 problem instances for different percentages of missing data, missing data patterns and noise levels σ. Best results are marked in bold.

	Missing data (%)	PCP [9]	WNNM [27]	Unifying [8]	LpSq [38]	S12L12 [46]	S23L23 [46]	IRNN [12]	APGL [47]	$\|\cdot\|_*$ [4]	$\mathcal{R}$ [35]	Our
Uniform ($\sigma = 0.0$)	0	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.0002	0.0002	**0.0000**	**0.0000**	0.1727	**0.0000**	**0.0000**
	10	0.0885	0.0028	0.0713	0.0213	0.0309	0.0071	**0.0000**	**0.0000**	0.1998	**0.0000**	**0.0000**
	20	0.2720	0.2220	0.1491	0.0170	0.0412	0.0209	**0.0000**	**0.0000**	0.2223	0.0128	**0.0000**
	30	0.7404	0.4787	0.7499	0.0003	0.0818	0.0895	**0.0000**	0.0014	0.2897	0.2346	**0.0000**
	40	1.0000	0.6097	0.9553	0.1083	0.1666	0.1360	**0.0000**	0.0017	0.3374	0.2198	**0.0000**
	50	1.0000	0.7170	1.0000	0.0315	0.1376	0.1001	0.0003	0.0301	0.4266	0.2930	**0.0000**
Tracking ($\sigma = 0.0$)	0	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.0002	0.0002	**0.0000**	**0.0000**	0.1810	**0.0000**	**0.0000**
	10	0.3160	0.2734	0.1534	0.0839	0.1296	0.1233	0.0772	0.0834	0.2193	0.0793	**0.0658**
	20	0.4877	0.4499	0.3017	0.1650	0.2389	0.2456	**0.1010**	0.1786	0.3436	0.2494	0.1018
	30	0.5821	0.5395	0.5486	0.2520	0.3289	0.3160	**0.1189**	0.2572	0.4299	0.3421	**0.1189**
	40	0.7072	0.6317	0.7376	0.2853	0.4084	0.4110	0.1417	0.2913	0.4825	0.5004	**0.1385**
	50	0.8125	0.7257	0.9521	0.4178	0.4267	0.4335	0.2466	0.4047	0.5754	0.6503	**0.2214**
Tracking ($\sigma = 0.1$)	0	0.0409	0.0207	0.0407	0.0450	0.0437	0.0435	0.0448	0.0191	0.1581	**0.0166**	**0.0166**
	10	0.3157	0.2734	0.1585	0.0848	0.0529	0.0518	0.0625	0.0696	0.2312	0.0488	**0.0438**
	20	0.4771	0.4338	0.3480	0.1394	0.0995	**0.0982**	0.1090	0.1188	0.3109	0.2071	0.0983
	30	0.5801	0.5225	0.4726	0.2026	0.2468	0.2592	0.1646	0.1993	0.3820	0.3465	**0.1475**
	40	0.7122	0.6148	0.8638	0.2225	0.3292	0.3252	0.1357	0.2110	0.4800	0.4599	**0.1273**
	50	0.7591	0.6819	0.9216	0.4105	0.4883	0.4811	0.3342	0.3639	0.5652	0.5930	**0.3329**

to the gradient w.r.t. X). In this section we show that by overparametrizing, in the sense that we use B and C with more columns than the rank of the solution we seek, it is still possible to use properties of (4) to show optimality in (7). We will exclusively use f_μ from (8), assume that B and C have $2k$ columns and study locally optimal solutions with $\operatorname{rank}(BC^T) < k$. The size of B and C makes it possible to parametrize line segments between such points and utilize convexity properties, see proof of Theorem 3. The following result (which is proven in the supplementary material) gives conditions that ensure that local minimality in (7) implies that (4) grows in all "low rank" directions.

Theorem 2. *Assume that* $(\bar{B}, \bar{C}) \in \mathbb{R}^{m\times 2k} \times \mathbb{R}^{n\times 2k}$, *where* $\bar{B} = U\sqrt{\Sigma}$ *and* $\bar{C} = V\sqrt{\Sigma}$, *and* $\bar{X} = U\Sigma V^T$, *is a local minimizer of* (7) *with* $\operatorname{rank}(\bar{X}) < k$ *and let* $\mathcal{N}(X) = \mathcal{R}(X) + \|\mathcal{A}X - b\|^2$. *Then* $\mathcal{R}(\bar{X}) = \tilde{\mathcal{R}}(\bar{B}, \bar{C})$ *and the directional derivatives* $\mathcal{N}'_{\Delta X}(\bar{X})$, *where* $\Delta X = \tilde{X} - \bar{X}$ *and* $\operatorname{rank}(\tilde{X}) \leq k$, *are non-negative.*

Note that there can be local minimizers for which $\tilde{\mathcal{R}}(\bar{B}, \bar{C}) > \mathcal{R}(\bar{B}\bar{C}^T)$ since $\tilde{\mathcal{R}}$ is non-convex. From an algorithmic point of view we can, however, escape such points by taking the current iterate and recompute the factorization of $\bar{B}\bar{C}^T$ using SVD. If the SVD of $\bar{B}\bar{C}^T = \sum_{i=1}^{r} \sigma_i U_i V_i^T$ we update $\bar{B}$ and $\bar{C}$ to $\bar{B}_i = \sqrt{\sigma_i} U_i$ and $\bar{C}_i = \sqrt{\sigma_i} V_i$, which we know reduces the energy and gives $\tilde{\mathcal{R}}(\bar{B}, \bar{C}) = \mathcal{R}(\bar{B}\bar{C}^T)$.

Theorem 2 allows us to derive optimality conditions using the properties of (4). As a simple example, consider the case where $\|\mathcal{A}X\|^2 \geq \|X\|^2$, which makes (4) convex [13], and let B and C have $2k$ columns. Suppose that we find a local minimizer $(\bar{B}, \bar{C})$ fulfilling the assumptions of Theorem 2. Then the derivative along a line segment towards any other low rank matrix is non-decreasing, and therefore $\bar{B}\bar{C}^T$ is the global optimum of (4) over the set of matrices with $\operatorname{rank} \leq k$ by convexity.

Below we give a result that goes beyond convexity and applies to the important class [45] of problems that obey the RIP constraint (3). Let $\mathcal{A}^*$ denote the adjoint operator of $\mathcal{A}$, then:

Theorem 3. *Assume that* $(\bar{B}, \bar{C})$ *is a local minimizer of* (7), *fulfilling the assumptions of Theorem 2. If the singular values of* $Z = (I - \mathcal{A}^*\mathcal{A})\bar{B}\bar{C}^T + \mathcal{A}^* b$ *fulfill* $\sigma_i(Z) \notin [(1 - \delta_{2k})\sqrt{\mu}, \frac{\sqrt{\mu}}{(1-\delta_{2k})}]$ *then* $\bar{B}\bar{C}^T$ *is the solution of* (9) *and* (10).

The proof builds on the results of [40] and is given in the supplementary material. The assumption that the singular values of Z are not too close to the threshold $\sqrt{\mu}$ is a natural restriction which is valid when the noise level is not too large. In case of exact data, *i.e.* $b = \mathcal{A}X_0$, where $\operatorname{rank}(X_0) = r$ it is trivially fulfilled for any choice of μ such that $\sqrt{\mu} < (1 - \delta_{2k})\sigma_r(X_0)$ since we have $Z = X_0$. For additional details on Z's dependence on noise see [14].

The above result is similar in spirit to those of [45, 28], which show that, in the convex case, having $2k$ columns and rank $2k - 1$ is enough to ensure that a local minimizer is global. For the proof in our non-convex case we need rank at most $k - 1$. Presently, it is not clear if our assumption can be relaxed to match that of the convex case or not.

4. An Iterative Reweighted VarPro Algorithm

In this section we give a brief overview of our algorithm for minimizing (7). A more detailed description is given in the supplementary material.

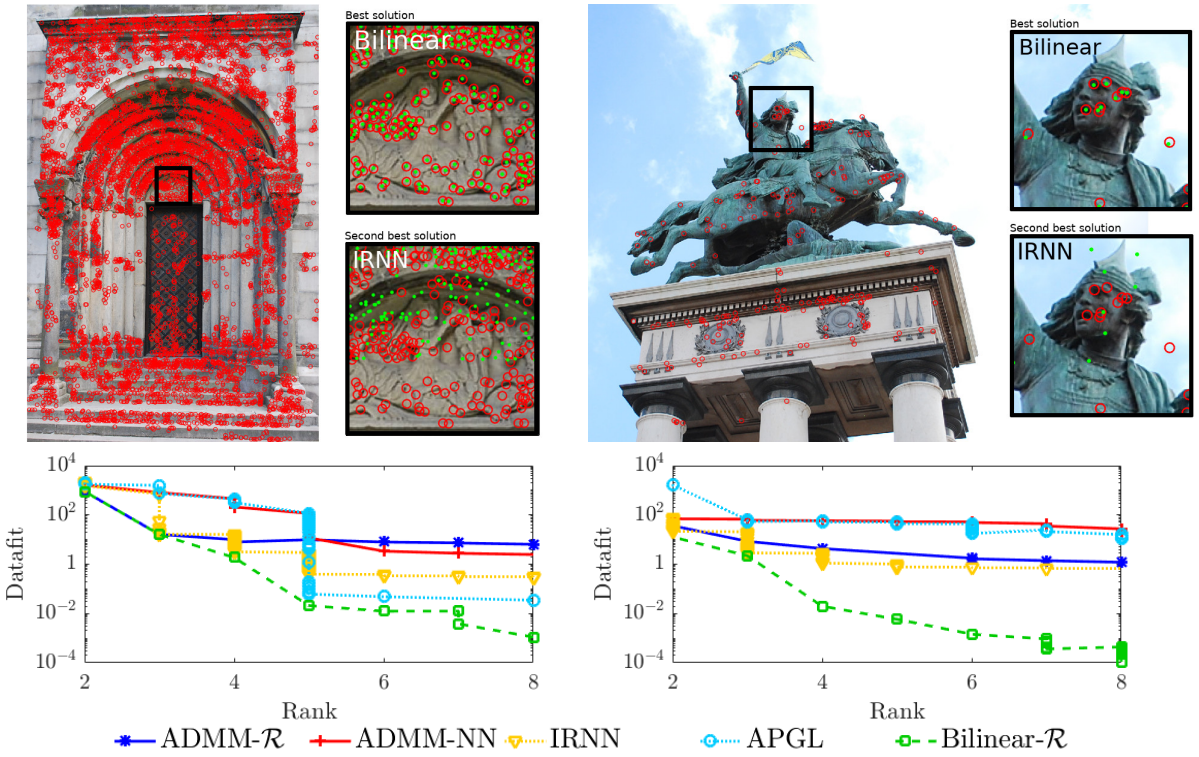

Figure 3. Comparison of reprojection error obtained using the bilinear formulation and ADMM, for datasets *Door* and *Vercingetorix* [41]. The red circles mark the feature points and the green dots the projected image points obtained from the different methods. The best rank 4 solution for the respective method was used. The control parameter $\eta = 0.5$ in both experiments.

Given a current iterate, $B^{(t)}$ and $C^{(t)}$, the first step of our algorithm is to replace the term $\tilde{\mathcal{R}}(B, C)$ with a quadratic function. To do this we note that by the Taylor expansion $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$, minimizing $f(x)$ and $f'(x_0)x$ around x_0 is roughly the same (ignoring constants). Inserting $x_0 = \frac{\|B_i^{(t)}\|^2 + \|C_i^{(t)}\|^2}{2}$ and $x = \frac{\|B_i\|^2 + \|C_i\|^2}{2}$ now gives our approximation

$$\sum_{i=1}^{k} w_i^{(t)} (\|B_i\|^2 + \|C_i\|^2) + \|\mathcal{A}BC^T - b\|^2, \qquad (20)$$

where $w_i^{(t)} = \frac{1}{2} f'\Big((\|B_i^{(t)}\|^2 + \|C_i^{(t)}\|^2)/2 \Big)$. Here $B_i^{(t)}$ and $C_i^{(t)}$ are the i:th columns of $B^{(t)}$ and $C^{(t)}$, respectively. Minimizing (20) over C is now a least squares problem with closed form solution. Inserting this solution into the original problem gives a nonlinear problem in B alone, which is what VarPro solves. We use the so called Ruhe and Wedin (RW2) approximation with a dampening term $\lambda \|B - B^{(t)}\|_F^2$, see [31] for details. In each step of the VarPro algorithm we update the weights $w_i^{(t)}$.

As previously mentioned, there can be stationary points for which $\tilde{\mathcal{R}}(B, C) > \mathcal{R}(BC^T)$. In each iteration we therefore take the current iterate and recompute the factorization of $B^{(t)}C^{(t)T}$ using SVD. If the SVD of $B^{(t)}C^{(t)T} = \sum_{i=1}^{r} \sigma_i U_i V_i^T$ we update $B^{(t)}$ and $C^{(t)}$ to $B_i^{(t)} = \sqrt{\sigma_i} U_i$ and $C_i^{(t)} = \sqrt{\sigma_i} V_i$ which we know reduces the energy and gives $\tilde{\mathcal{R}}(B^{(t)}, C^{(t)}) = \mathcal{R}(B^{(t)}C^{(t)T})$.

Our approach can be seen as iteratively reweighted nuclear norm minimization [12]; however, our bilinear formulation allows us to use quadratic approximation, thus benefiting from second order convergence in the neighborhood of a local minimum.

5. Experiments

In this section we will show the versatility and strength of the proposed method, focusing on computer vision problems. In Section 5.2 we show an example where state-of-the-art methods fail to achieve a value close to global optimality. We include two more examples of real problems, in the supplementary material: background extraction and photometric stereo. In both cases our method shows superior performance. In the main paper we focus on the trade-off between datafit and rank, but show, in the examples in the supplementray material, the added benefits of convergence speed using the proposed method. This is done by minimizing the same energy with ADMM and the proposed method, in which case the splitting schemes can be tediously slow. In all experiments our proposed method is initialized randomly, with zero mean and unit variance.

5.1. Synthetic Missing Data Problem

Let $\odot$ denote the Hadamard product, and consider the missing data formulation

$$\min_X \mu \text{rank}(X) + \|W \odot (X - M)\|_F^2 , \qquad (21)$$

where M is a measurement matrix and W a missing data mask with entries $w_{ij} = 1$ if the entry is known, and zero otherwise.

In low-level vision applications such as denoising and image inpainting a uniformly random missing data pattern is often a reasonable approximation of the distribution; however, for structure from motion, the missing data pattern is often highly structured. To this end, we investigate two kinds of patterns: uniformly random and "tracking failure". In order to construct realistic patterns of tracking failure, we use the method in [36]. This is done by randomly selecting if a track should have missing data (with uniform probability), then select (with uniform probability, starting after the first few frames) in which image tracking failure occurs. If a track is lost, it is not restarted.

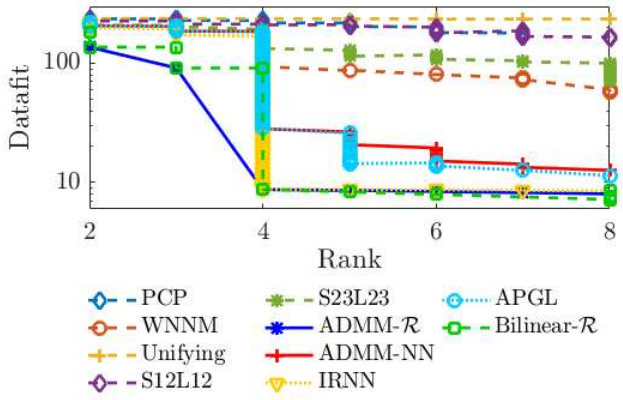

Figure 4. Rank vs datafit for the synthetic experiment in Section 5.1. No true low rank solution using LpSq [38] could be found, regardless of the choice of parameters.

We generate random ground truth matrices $M_0 \in \mathbb{R}^{32\times512}$ of rank 4, which can be expressed as $M_0 = UV^T$, where $U \in \mathbb{R}^{32\times4}$ and $V \in \mathbb{R}^{512\times4}$. The entries of U and V are normal distributed with zero mean and unit variance. The measurement matrix $M = M_0 + N$, where N simulates noise and has normal distributed entries with zero mean and variance σ^2.

Our proposed method is compared to a variety of different methods [8, 9, 27, 38, 46, 12, 47, 4, 35]. For the methods that need an initial estimate of the rank as input, the rank estimation heuristic by Shang *et al.* [46] is used. The regularization parameter is set to $\lambda = \sqrt{\max(m, n)}$, given a sought $m \times n$ matrix, as proposed by [9, 46]. In case other parameters should be provided, the one recommended from the respective authors have been used. The number of columns, for our proposed method, is set to $k = 8$, *i.e.* twice the rank of the original matrix M_0. We exclusively use the f_μ regularization (8), and use $\sqrt{\mu} = \lambda$. Since f_μ is a special case of MCP, it is used for IRNN as well. Furthermore, we include the results for regularizing with nuclear norm [4] and f_μ (8) using ADMM, as proposed in [35]. Note that ADMM comes without optimality guarantees, however, it has been shown to work well for several computer vision problems in practice [35, 40]. Several of the compared methods solve the robust PCA problem, thus also include a sparse component, which is not taken into account.

The results are shown in Table 1. Note that most algorithms perform significantly better for the uniformly random missing data pattern, than compared to the structured missing data pattern. Our proposed method outperforms all other methods in this comparison.

Since the final rank of the estimated matrix is not necessarily the same as that of M_0, we show the rank vs datafit obtained when varying the regularization parameter λ in Figure 4. It is evident from the results that the only candidates that yield an acceptable result for low rank solutions are ADMM with f_μ, IRNN with MCP and our proposed method.

5.2. pOSE: Pseudo Object Space Error

The Pseudo Object Space Error (pOSE) objective combines affine and projective camera models

$$\ell_{\text{OSE}} = \sum_{(i,j)\in\Omega} \left\| (P_{i,1:2}\tilde{\mathbf{x}}_j - (\mathbf{p}_{i,3}^T\tilde{\mathbf{x}}_j)\mathbf{m}_{i,j}) \right\|^2 , \qquad (22)$$

$$\ell_{\text{Affine}} = \sum_{(i,j)\in\Omega} \left\| P_{i,1:2}\tilde{\mathbf{x}}_j - \mathbf{m}_{i,j} \right\|^2 , \qquad (23)$$

$$\ell_{\text{pOSE}} = (1-\eta)\ell_{\text{OSE}} + \eta\ell_{\text{Affine}}, \qquad (24)$$

where ℓ_{OSE} is the object space error and ℓ_{Affine} is the affine projection error. Here $P_{i,1:2}$ denotes the first two rows, $\mathbf{p}_{i,3}$ the third row of the i:th camera matrix, and $\tilde{\mathbf{x}}_j$ is the j:th 3D point in homogeneous coordinates. The control parameter $\eta \in [0, 1]$ determines the impact of the respective camera model. This objective was introduced in [33] to be used in a first stage of an initialization-free bundle adjustment pipeline, optimized using VarPro.

The ℓ_{pOSE} objective is linear, and acts on low-rank components P and X, which are constrained by $\text{rank}(PX^T) = 4$. Instead of enforcing the rank constraint, we replace it as before with a relaxation. By not enforcing the rank constraint we demonstrate the ability of the methods to make accurate trade-offs between minimizing the rank and fitting the data. Since the objective now becomes more complex, and is no longer compatible with the missing data formulations, only IRNN and APGL are directly applicable, as well as the ADMM approach using f_μ

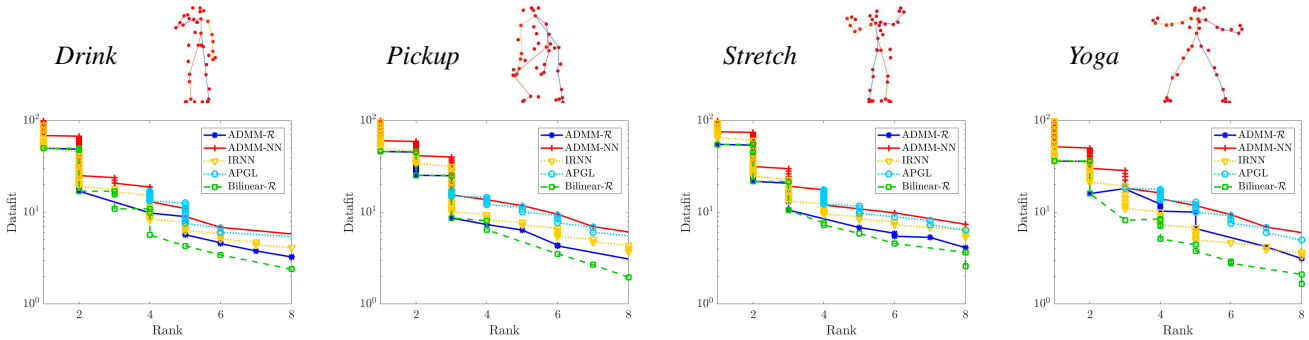

Figure 5. *Top row:* Example frames from the MOCAP dataset of the *drink*, *pickup*, *stretch* and *yoga* sequences. *Last row:* The bilinear method finds the same or a better datafit compared to the other methods for all ranks.

and nuclear norm. We use two real-life datasets with various amounts of camera locations and 3D points: *Door* with 12 images, resulting in seeking a matrix of size 36×8850 and *Vercingetorix* [41] with 69 images, resulting in seeking a matrix of size 207×1148, both of which have rank 4. [2]

As in the synthetic experiment from Section 5.1, the regularization parameter is varied and the resulting rank and datafit is stored and reported in Figure 3. To visualize the results, we considered the best rank 4 approximations, and show the reprojected points and the corresponding measured points obtained from the best method (ours in both cases) and the second best (IRNN in both cases), see Figure 3. As is readily seen by ocular inspection, the rank 4 solution obtained by our proposed method significantly outperforms those of other state-of-the-art methods.

5.3. Non-Rigid Structure From Motion

In this section we test our approach on non-rigid reconstruction (NRSfM) with the CMU Motion Capture (MOCAP) dataset. In NRSfM, the complexity of the deformations are controlled by some mild assumptions of the object shapes. Bregler *et al*. [6] suggested that the set of all possible configurations of the objects are spanned by a low dimensional linear basis of dimension K. In this setting, the non-rigid shapes $X_i \in \mathbb{R}^{3\times n}$ can be represented as $X_i = \sum_{k=1}^{K} c_{ik} B_k$, where $B_k \in \mathbb{R}^{3\times n}$ are the basis shapes and $c_{ik} \in \mathbb{R}$ the shape coefficients. This way, the matrix X_i contains the world coordinates of point i, hence the observed image points are given by $x_i = R_i X_i$. We will assume orthographic cameras, *i.e.* $R_i \in \mathbb{R}^{2\times 3}$ where $R_i R_i^T = I_2$. As proposed by Dai *et al*. [15], the problem can be turned into a low-rank factorization problem by reshaping and stacking the non-rigid shapes X_i. Let $X_i^\sharp \in \mathbb{R}^{1\times 3n}$ denote the concatenation of the rows in X_i, and create $X^\sharp \in \mathbb{R}^{F\times 3n}$ by stacking $X_i^\sharp$. This allows us to decompose the matrix $X^\sharp$ in the low-rank factors $X^\sharp = CB^\sharp$, where $C \in \mathbb{R}^{F\times K}$ contains the shape coefficients c_{ik} and $B^\sharp \in \mathbb{R}^{K\times 3n}$ is constructed as $X^\sharp$ and contains the basis elements.

A suitable objective function is thus given by

$$\mu \text{rank}(X^\sharp) + \|RX - M\|_F^2 , \tag{25}$$

where $R \in \mathbb{R}^{2F\times 3F}$ is a block-diagonal matrix with the camera matrices R_i on the main diagonal, $X \in \mathbb{R}^{3F\times n}$ is the concatenation of the 3D points X_i, and $M \in \mathbb{R}^{2F\times n}$ is the concatenated observed image points x_i. By replacing the rank penalty with a relaxation and minimize it using the proposed method and the methods used in the previous section. The regularization parameter is varied for the respective methods in order to obtain a rank 1–8 solution, and the respective datafit is reported in Figure 5, for four different sequences.

In all sequences, the best datafit for each rank level is obtained by our proposed method. IRNN and ADMM using f_μ is able to give the same, or very similar, datafit for lower ranks, but for solutions with rank larger than four our method consistently reports a lower value than the competing state-of-the-art methods.

6. Conclusions

In this paper we presented a unification of bilinear parameterization and rank regularization. Robust penalties for rank regularization has often been used together with splitting schemes, but it has been shown that such methods yield unsatisfactory results for ill-posed problems in several computer vision applications. By using the bilinear formulation, the objective functions become differentiable, and convergence rates in the neighborhood of a local minimum are faster. Furthermore, we showed that theoretical optimality results known from the regularization formulations can be lifted to the bilinear formulation.

Lastly, the generality of the proposed framework allows for a wide range of problems, some of which, have not been amenable by state-of-the-art methods, but have been proven successful using our proposed method.

[2] The datasets are available here: http://www.maths.lth.se/matematiklth/personal/calle/dataset/dataset.html.

References

[1] F. R. Bach. Convex relaxations of structured matrix factorizations. *CoRR*, abs/1309.3117, 2013. 2, 3

[2] R. Basri, D. Jacobs, and I. Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3):239–257, May 2007. 1

[3] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Annual Conference in Neural Information Processing Systems (NIPS)*. 2016. 1

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011. 2, 5, 7

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 2

[6] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000. 1, 8

[7] A. M. Buchanan and A. W. Fitzgibbon. Damped newton algorithms for matrix factorization with missing data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 1

[8] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *International Conference on Computer Vision (ICCV)*, 2013. 2, 3, 5, 7

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011. 2, 5, 7

[10] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. 2

[11] L. Canyi, J. Tang, S. Yan, and Z. Lin. Generalized nonconvex nonsmooth low-rank minimization. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[12] L. Canyi, J. Tang, S. Yan, and Z. Lin. Nonconvex nonsmooth low-rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25, 10 2015. 2, 5, 6, 7

[13] M. Carlsson. On convexification/optimization of functionals including an l2-misfit term. *arXiv preprint arXiv:1609.09378*, 2016. 4, 5

[14] M. Carlsson, D. Gerosa, and C. Olsson. An unbiased approach to compressed sensing. *arXiv preprint*, arXiv:1806.05283, 2018. 5

[15] Y. Dai, H. Li, and M. He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 8

[16] A. Eriksson and A. Hengel. Efficient computation of robust weighted low-rank matrix approximations using the L_1 norm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1681–1690, 2012. 1

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. 3

[18] M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, 2001. 2

[19] J. H. Friedman. Fast sparse regression and classification. *International Journal of Forecasting*, 28(3):722 – 738, 2012. 3

[20] C. Gao, N. Wang, Q. R. Yu, and Z. Zhang. A feasible nonconvex relaxation approach to feature selection. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2011. 3

[21] R. Garg, A. Roussos, and L. Agapito. A variational approach to video registration with subspace constraints. *International Journal of Computer Vision*, 104(3):286–314, 2013. 1

[22] R. Garg, A. Roussos, and L. de Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[23] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint*, arxiv:1704.00708, 2017. 1

[24] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016. 1

[25] D. Geman and C. Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4:932 – 946, 08 1995. 3

[26] N. Gillis and F. Glinuer. Low-rank matrix approximation with weights or missing data is np-hard. *SIAM Journal on Matrix Analysis and Applications*, 32(4), 2011. 1

[27] S. Gu, Q. Xie, D. Meng, W. Zuo, X. Feng, and L. Zhang. Weighted nuclear norm minimization and its applications to low level vision. *International Journal of Computer Vision*, 121, 07 2016. 2, 5, 7

[28] B. D. Haeffele and R. Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 3, 5

[29] J. H. Hong and A. Fitzgibbon. Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 3

[30] J. H. Hong, C. Zach, A. Fitzgibbon, and R. Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *European Conference on Computer Vision (ECCV)*, 2016. 1, 3

[31] J. H. Hong, C. Zach, A. Fitzgibbon, and R. Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3, 6

[32] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2117–2130, 2013. 2

[33] J. Hyeong Hong and C. Zach. pose: Pseudo object space error for initialization-free bundle adjustment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 3, 7

[34] M. Krechetov, J. Marecek, Y. Maximov, and M. Takac. Entropy-penalized semidefinite programming. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1123–1129. International Joint Conferences on Artificial Intelligence Organization, 7 2019. 3

[35] V. Larsson and C. Olsson. Convex low rank approximation. *International Journal of Computer Vision*, 120(2):194–214, 2016. 2, 3, 4, 5, 7

[36] V. Larsson and C. Olsson. Compact matrix factorization with dependent subspaces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4361–4370, 07 2017. 7

[37] K. Mohan and M. Fazel. Iterative reweighted least squares for matrix rank minimization. In *Annual Allerton Conference on Communication, Control, and Computing*, pages 653–661, 2010. 2

[38] F. Nie, H. Wang, X. Cai, H. Huang, and C. H. Q. Ding. Robust matrix completion via joint schatten p-norm and lp-norm minimization. In *ICDM*, pages 566–574, 2012. 2, 5, 7

[39] 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 Transactions on Pattern Analysis and Machine Intelligence*, 38(4):744–758, 2016. 2

[40] C. Olsson, M. Carlsson, F. Andersson, and V. Larsson. Non-convex rank/sparsity regularization and local minima. *Proceedings of the International Conference on Computer Vision*, 2017. 2, 5, 7

[41] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, pages 524–535, 2011. 6, 8

[42] M. V. Örnhag, C. Olsson, and A. Heyden. Differentiable fixed-rank regularisationusing bilinear parameterisation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 3

[43] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5):2886–2908, 2015. 2

[44] S. Oymak, K. Mohan, M. Fazel, and B. Hassibi. A simplified approach to recovery conditions for low rank matrices. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 2318–2322, 2011. 2

[45] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, Aug. 2010. 1, 2, 5

[46] F. Shang, J. Cheng, Y. Liu, Z. Luo, and Z. Lin. Bilinear factor matrix norm minimization for robust pca: Algorithms and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9):2066–2080, Sep. 2018. 2, 3, 4, 5, 7

[47] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6, 09 2010. 2, 5, 7

[48] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 1

[49] M. Uchiyama. Subadditivity of eigenvalue sums. *Proceedings of The American Mathematical Society*, 134:1405–1412, 05 2005. 4

[50] N. Wang, T. Yao, J. Wang, and D.-Y. Yeung. A probabilistic approach to robust matrix factorization. In *European Conferencec on Computer Vision (ECCV)*, 2012. 1

[51] C. Xu, Z. Lin, and H. Zha. A unified convex surrogate for the schatten-p norm. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2017. 3

[52] J. Yan and M. Pollefeys. A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):865–877, 2008. 1

[53] C.-H. Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. 3, 4