Bilinear Parameterization for Non-Separable Singular Value Penalties

Marcus Valtonen Örnhag[1] José Pedro Iglesias[2] Carl Olsson[1,2]

[1]Centre for Mathematical Sciences
Lund University

[2]Department of Electrical Engineering
Chalmers University of Technology

Abstract

Low rank inducing penalties have been proven to successfully uncover fundamental structures considered in computer vision and machine learning; however, such methods generally lead to non-convex optimization problems. Since the resulting objective is non-convex one often resorts to using standard splitting schemes such as Alternating Direction Methods of Multipliers (ADMM), or other subgradient methods, which exhibit slow convergence in the neighbourhood of a local minimum. We propose a method using second order methods, in particular the variable projection method (VarPro), by replacing the non-convex penalties with a surrogate capable of converting the original objectives to differentiable equivalents. In this way we benefit from faster convergence.

The bilinear framework is compatible with a large family of regularizers, and we demonstrate the benefits of our approach on real datasets for rigid and non-rigid structure from motion. The qualitative difference in reconstructions show that many popular non-convex objectives enjoy an advantage in transitioning to the proposed framework.[1]

1. Introduction

Low rank approximation and factorization methods are classical approaches for solving various computer vision problems, such as structure from motion [38, 5, 21, 11, 24, 25], photometric stereo [2, 7, 29] image segmentation [15], image restoration [9, 29, 20, 27, 45], background/foreground segmentation [43, 7], etc.

There are two main approaches when it comes to solving these problems, and which one is used largely depends on properties of the particular problem being addressed. The classical problem of low rank recovery with missing data

$$\min_{\text{rank}(X)\leq k} \|W \odot (X - M)\|_F^2, \tag{1}$$

is a core step in many structure from motion formulations [6]. Here M is a measurement matrix which is only partially known and W is a binary matrix removing residuals corresponding to unknown elements. The traditional approach, which is typically used when the rank of the sought matrix is known, enforces a particular rank by restricting the number of columns of the factors B and C and searches over the bilinear parameterization of the unknown matrix $X = BC^T$. Since the resulting objective is a least squares problem in both B and C, alternating updates of B and C can be used. While being extremely simple, this approach has been shown to be prone to "flatlining: requiring excessive numbers of iterations before convergence" [6]. Instead [6] proposed a damped newton approach and empirically

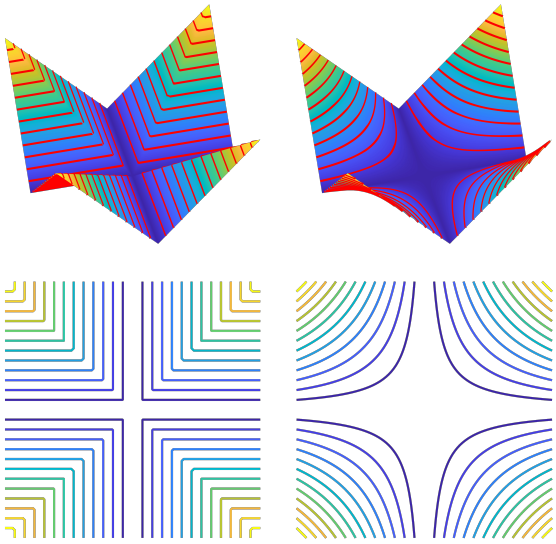


Figure 1: *Top row:* Weighted nuclear norm penalty (left) and the corresponding relaxation r_h, for $a_1 = 0$ and $a_2 = 1$, considered in this paper. *Bottom row:* Level sets, corresponding to the red lines in the top images.

[1]This work was supported by the Swedish Research Council (grants no. 2015-05639, 2016-04445 and 2018-05375), the strategic research project ELLIIT, the Swedish Foundation for Strategic Research (Semantic Mapping and Visual Navigation for Smart Robots) and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

verified that this outperforms the alternation approach. In a number of recent papers Hong *et al*. [17, 19, 18, 21] showed that the so called VarPro method is remarkably resilient to local minima. For example [18] reports convergence to the best solution from random initialization in 94% of the cases on the dinosaur sequence which is an admittedly difficult dataset with 77% missing data.

An alternative approach is to optimize directly over the elements of X while applying penalties to the singular values. This is typically applied to problems of the more general class

$$\min_X p(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - b\|^2. \tag{2}$$

Here p is some penalty function encouraging a desired distribution of singular values $\sigma_i(X)$ of the matrix X, see *e.g.* [29, 20, 27]. This way of directly optimizing over the elements of X has been made popular by the work on nuclear norms [34, 8] and their generalizations [26, 28, 12, 14, 13], which has shown that with an appropriate choice of regularizer (2) can be made convex. While convex regularizers can be sufficient for applications such as image restoration, where a relatively high rank is acceptable, they are typically rather weak and do not give solutions with low enough rank for structure from motion problems. Consequently, they have to be combined with thresholding schemes to generate satisfactory solutions [7, 11].

To achieve better results, non-convex penalty functions [25, 29, 16, 9] are also frequently used in (2). Since these formulations are typically not differentiable, optimization relies on splitting methods such as ADMM [4]. These are essentially first order methods and, as such, convergence near the minimum can be slow. Indeed, [4] recommends to use these when an approximate solution is sufficient, but suggests to switch to second order methods when accuracy is needed.

In this paper, we derive such second order methods for a general class of objectives of the form (2). Our class covers commonly used regularizers, such as weighted nuclear norms, soft rank penalties and hard rank constraints. Note that these functions can be both non-convex and discontinuous. We show how to reformulate these into bilinear objectives, that can be accurately approximated with quadratic functions, allowing rapid convergence with second order methods such as VarPro or Levenberg–Marquardt.

1.1. Framework and Contributions

In this paper we consider a general framework of non-separable objectives the form

$$f_h(X) = h(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - b\|^2, \tag{3}$$

where

$$h(\boldsymbol{\sigma}(X)) = \sum_{i=1}^{\operatorname{rank}(X)} a_i\sigma_i(X) + b_i. \tag{4}$$

Here the sequences $(a_i)_{i=1}^k$ and $(b_i)_{i=1}^k$ are both assumed to be non-decreasing. For different choices of a and b the general regularizer h reduces to commonly used singular value penalties. For example, with $b_i = \mu$ and $a_i \equiv 0$ we get the soft rank penalty $\mu\operatorname{rank}(X)$, with $a_i \equiv 0$ if $i \leq k$ and ∞ otherwise, we get the hard constraint $\operatorname{rank}(X) \leq k$. With $b_i \equiv 0$, we get the weighted nuclear norm $\sum_i a_i\sigma_i(X)$, but the framework is large and several other regularizers are possible. We aim to optimize objectives including all such regularizers using second order methods, such as VarPro. This requires finding a good approximation—which is two times differentiable—of the objective function. For this purpose, we propose to use a relaxation $r_h(\boldsymbol{\sigma})$ of $h(\boldsymbol{\sigma})$ developed in [40] and consider

$$r_h(\boldsymbol{\sigma}(X)) + \|\mathcal{A}X - b\|^2. \tag{5}$$

This results in a continuous and almost everywhere differentiable objective (see Section 2 for details, Figures 1 and 2 for examples). When introducing the terms $\gamma_i(B, C) = (\|B_i\|^2 + \|C_i\|^2)/2$, where B_i and C_i are columns i of B and C, respectively, we obtain the bilinear formulation

$$\min_{B,C} r_h(\boldsymbol{\gamma}(B, C)) + \|\mathcal{A}(BC^T) - b\|^2. \tag{6}$$

The main contributions of this paper are:

i) We show that (5) is equivalent to (6) by proving that

$$r_h(\boldsymbol{\sigma}(X)) = \min_{X=BC^T} r_h(\boldsymbol{\gamma}(B, C)), \tag{7}$$

see Theorem 1. Furthermore, if $\|\mathcal{A}\| < 1$ then the relaxation (5) is guaranteed to have the same global optimizers as the original (3), see [10].

ii) We show that (6) can be accurately approximated by quadratic functions opening up the possibility of applying second order methods to the problem.

iii) We propose a modified VarPro algorithm and show that it provides superior performance for difficult objectives common in computer vision applications.

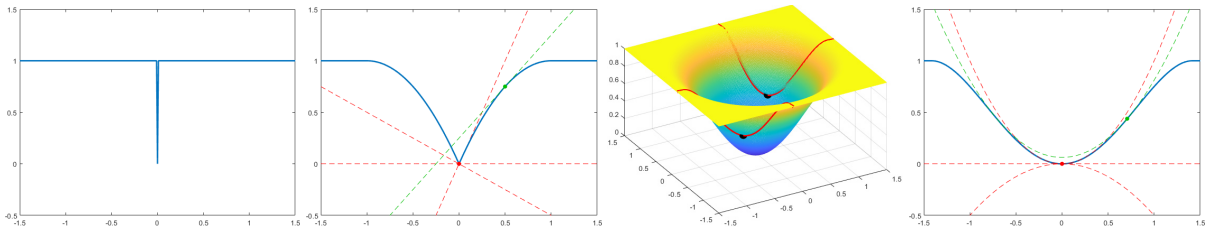
1.2. Related Work

There are some works that have previously proposed bilinear formulations for problems of the form (2). It was observed in [34] that

$$\|X\|_* = \min_{BC^T=X} \frac{\|B\|_F^2 + \|C\|_F^2}{2}. \tag{8}$$

Thus, when $p(\boldsymbol{\sigma}(X)) = \|X\|_*$ optimization of (2) can be formulated as

$$\min_{B,C} \mu\frac{\|B\|_F^2 + \|C\|_F^2}{2} + \|\mathcal{A}BC^T - b\|^2. \tag{9}$$



(a) (b) (c) (d)

Figure 2: (a): One dimensional example of $h(x)$, with $a_1 = 0$ and $b_1 = 1$. (b): The relaxation $r_h(x)$. Replacing h with r_h makes the objective continuous and differentiable everywhere except in $x = 0$. In this point r_h has several subgradients (red lines) while on any other point has it unique tangent (e.g. green line). (b): The two dimensional function $r_h(\frac{b^2+c^2}{2})$. Parametrizing with squared variables smooths the function around around $(b, c) = (0.0)$ giving an objective that can be locally well approximated with quadratic functions. (d): The function from (c), sliced along $c = 0$, with the quadratic approximations resulting from the subgradients shown in (b).

Interestingly, even though (9) is non-convex, it can be shown that any local minimizer B,C with $\text{rank}(BC^T) < k$, where k is the number of columns in B and C, is globally optimal [1, 15]. In addition, the objective function is smooth and second order methods can be employed.

A similar formulation was tested for structure from motion and photometric stereo in [7]. In practice, it was, however, observed that the nuclear norm is too weak to give low rank solutions when the data is noisy. Therefore, a "continuation" approach, where the size of the factorization is gradually reduced, is proposed. In [15] the above results where extended to a general class of 2-homogeneous factor penalties $\theta(B, C)$. Interestingly, this formulation allows to add constraints such as non-negativity to the factors B and C. Their results show that $\min_{BC^T=X} \theta(B, C)$ is equivalent to a convex regularization function $\Omega(X)$. The property that a local minimizer B, C with $\text{rank}(BC^T) < k$ is global, is also extended to this class of functions.

Similar approaches to non-convex formulations have also been proposed. Shang *et al.* [36] 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 other values of q is given in [44]. A separable regularizer of the form $p(\boldsymbol{\sigma}(X)) = \sum_{i=1}^{n} \tilde{p}(\boldsymbol{\sigma}_i(X))$ was considered in [41]. It was shown that when $\tilde{p}(\sigma)$ is concave and non-decreasing on $\sigma \geq 0$ then

$$p(\boldsymbol{\sigma}(X)) = \min_{BC^T=X} \sum_{i=1}^{n} \tilde{p}(\gamma_i(B, C)). \tag{10}$$

Separable penalties, such as the once mentioned above, are limited in the sense that they only consider the singular values separately. The penalty can therefore only be based on the magnitude of the singular values. As a consequence, they cannot count the number of non-zero singular values, thus making it impossible to penalize all matrices with rank larger than a predefined threshold. A non-separable formulation has very recently been addressed in [22]. Here (10) was generalized to weighted nuclear norms, that is $p(\boldsymbol{\sigma}(X)) = \sum_{i=1}^{n} a_i \sigma_i(X)$. Note that, in contrast to the original nuclear norm, these are not convex since they are able to penalize smaller singular values harder (when $(a_i)_{i=1}^n$ is increasing).

One of the main benefits of our framework is that there are provably fewer local minima [10], compared to using the unrelaxed penalty used in [22]. Valtonen Örnhag *et al.* [33] considered a special case of our framework, where the rank is known a priori. They did, however, not show equivalence between the proposed bilinear regularizer and the corresponding original formulation.

2. Overview of the Approach

In this section we present our general framework and explain and motivate our algorithmic approach.

2.1. A Continuous Relaxation

Objective (3) is typically discontinuous, and, therefore, local approximation (linear or quadratic) is not directly feasible. To circumvent this issue, we consider the quadratic envelop [10] $r_h(\mathbf{x})$ of $h(\mathbf{x})$

$$r_h(\mathbf{x}) = \left(h(\mathbf{x}) + \|\mathbf{x}\|^2\right)^{**} - \|\mathbf{x}\|^2, \tag{11}$$

that is, we add a quadratic term to h, compute the convex envelope of the result, and subtract the quadratic term. Throughout the paper we will use the function $g(\mathbf{x}) \coloneqq r_h(\mathbf{x}) - \|\mathbf{x}\|^2 = \left(h(\mathbf{x}) + \|\mathbf{x}\|^2\right)^{**}$. In general, r_h and g do not have closed form expressions, but are obtained from

$$g(\mathbf{x}) = \max_{\mathbf{z} \in \mathbb{R}^n} \left(2\langle \mathbf{x}, \mathbf{z} \rangle - \sum_{i=1}^{n} \left[\left[|z_{[i]}| - a_i \right]_+^2 - b_i \right]_+ \right), \tag{12}$$

where $[\cdot]_+ := \max(\cdot, 0)$ and $z_{[i]}$ denotes the element with the i:th largest magnitude in $\mathbf{z}$. The optimization over $\mathbf{z}$ is convex and can be solved efficiently, as outlined in [40].

Note that in (3) the function h always takes sorted non-negative vectors $\boldsymbol{\sigma}(X)$. For a general vector $\mathbf{x}$ we therefore think of h as being permutation and sign invariant, that is

$$h(\mathbf{x}) = \sum_{i=1}^{\text{card}(\mathbf{x})} a_i \left|x_{[i]}\right| + b_i, \tag{13}$$

where $\text{card}(\mathbf{x})$ is the number of non-zero elements in $\mathbf{x}$. Figure 1 shows an example of r_h and h with $b_i = 0$, *i.e.* $h(\boldsymbol{\sigma}(X))$ is a weighted nuclear norm of X. Both functions are permutation invariant; however, r_h is additionally continuous and differentiable almost everywhere, except where one of the variables is equal to zero. Figure 2a shows a one dimensional example where $a_1 = 0$ and $b_1 = 1$. In this case the function h is discontinuous, which is generally the case when any of the b_i variables are non-zero. Figure 2b shows the relaxation r_h for the choice of h in 2a.

Replacing h with r_h gives us the relaxation (5) which in terms of regularity is significantly better behaved than (3). There are, of course, many ways of approximating the h function. The reason for choosing this particular relaxation is that it can be shown [10] that if $\|\mathcal{A}\| < 1$ it has the same global minimizer as (3). In addition it has been shown that if RIP [34] holds, then, under moderate noise, (5) only has one stationary point for some choices of h [30, 31].

2.2. Bilinear Parameterization

The regularity of our formulation improves even further when we introduce the bilinear terms $\gamma_i(B, C) = (\|B_i\|^2 + \|C_i\|^2)/2$ and consider (6). Here we have replaced X with a factorization BC^T and instead of penalizing the singular values of X we penalize the vector $\boldsymbol{\gamma}(B, C)$ containing the elements $\gamma_i(B, C)$. Note that the regularizer $r_h(\boldsymbol{\gamma}(B, C))$ depends on which particular factorization $BC^T = X$ of X that we chose. The main theoretical result of this paper states that for any fixed matrix X we have

$$\min_{BC^T = X} r_h(\boldsymbol{\gamma}(B, C)) = r_h(\boldsymbol{\sigma}(X)), \tag{14}$$

hence, minimization over factorizations BC^T of X will result in the same penalty as the singular value vector $\boldsymbol{\sigma}(X)$. Furthermore, since the second term $\|\mathcal{A}(BC^T) - b\|^2 = \|\mathcal{A}X - b\|^2$, regardless of which factorization we choose, it is clear that minimization of (6) is equivalent to (5).

Figure 2c shows $r_h(\frac{b^2+c^2}{2})$ as a function of (b, c). Note that introducing the squared variables makes the resulting function smooth at $(b, c) = (0, 0)$. To illustrate our main result, we also plot red curves on the surface corresponding to all points where $bc = 0.25$. While the regularization term can take many values over $bc = 0.25$, the two minimizers $(b, c) = \pm(0.5, 0.5)$, shown as black dots, both give the value $r_h\left(\frac{0.5^2+0.5^2}{2}\right) = r_h(0.25)$.

2.3. Quadratic Approximation and Optimization

The basis for our algorithm is the ability to accurately approximate $r_h(\boldsymbol{\gamma}(B, C))$ with a quadratic relaxation. The principle can be illustrated by considering a general smooth function $r(\boldsymbol{\gamma})$ with the first order Taylor approximation around $\boldsymbol{\eta}$

$$r(\boldsymbol{\gamma}) \approx r(\boldsymbol{\eta}) + \langle \nabla r(\boldsymbol{\eta}), \boldsymbol{\gamma} - \boldsymbol{\eta} \rangle. \tag{15}$$

When inserting $\boldsymbol{\gamma} = \boldsymbol{\gamma}(B, C)$ we obtain (ignoring the constants $r(\boldsymbol{\eta})$ and $-\langle \nabla r(\boldsymbol{\eta}), \boldsymbol{\eta} \rangle$) the quadratic approximation

$$\langle \nabla r(\boldsymbol{\eta}), \boldsymbol{\gamma}(B, C) \rangle = \sum_{i=1}^{n} r_i'(\boldsymbol{\eta}) \frac{\|B_i\|^2 + \|C_i\|^2}{2}, \tag{16}$$

where r_i' is the partial derivative with respect to the i:th entry of r. Since our particular regularizer r_h is not differentiable everywhere (when parametrized with $\mathbf{x}$) we make use of the so called subdifferential $\partial r_h(\mathbf{x})$ of r_h. For a convex function g we have $2\mathbf{z} \in \partial g(\mathbf{x})$ if and only if

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \langle 2\mathbf{z}, \mathbf{y} - \mathbf{x} \rangle, \tag{17}$$

for all $\mathbf{y}$. For the g defined in (12) it can be shown that (17) holds if and only if $\mathbf{z}$ is a minimizer in (12). At a point where the function is differentiable the gradient is the only element in the subdifferential, and the right hand side of (17) is the Taylor approximation. However, in general it can contain several vectors and these can be seen as lower bounding linear approximations, as in (17).

If, as in our case, $r_h(\mathbf{x}) = g(\mathbf{x}) - \|\mathbf{x}\|^2$ we have that $2(\mathbf{z} - \mathbf{x}) \in \partial r_h(\mathbf{x})$ when $2\mathbf{z} \in \partial g(\mathbf{x})$. For a non-convex function the above inequality becomes approximate around $\mathbf{x}$ (up to higher order terms). Figure 2b shows examples of subgradients at two points. At $x = 0$ (red point) where the function is not differentiable there are several such linear bounds, while at $x = 0.5$ there is only one tangent. For our class of functions, an element of z_i is only non-unique when $x_i = 0$.

Suppose now that we want to approximate $r_h(\boldsymbol{\gamma}(B, C))$ around a point $\boldsymbol{\eta} = \boldsymbol{\gamma}(\bar{B}, \bar{C})$. Up to higher order terms, we then have

$$r_h(\boldsymbol{\gamma}) \geq r_h(\boldsymbol{\eta}) - 2\langle \mathbf{z} - \boldsymbol{\eta}, \boldsymbol{\eta} \rangle + 2\sum_{i=1}^{n} (z_i - \eta_i) \frac{\|B_i\|^2 + \|C_i\|^2}{2}. \tag{18}$$

Since $2\mathbf{z} \in \partial g(\boldsymbol{\eta})$ it can be shown that z_i can only take one value when $\eta_i \neq 0$, see also Figure 2 (we also give a proof of this in the supplementary material). Therefore the terms $r_h(\boldsymbol{\eta})$ and $\langle \mathbf{z} - \boldsymbol{\eta}, \boldsymbol{\eta} \rangle$ are constants. In addition, it can be shown that for any i where $\eta_i = 0$, z_i can be chosen freely as long as its magnitude is smaller than a certain number

Table 1: Log-loss for the Door, Back, Heart, and Paper datasets, for $\eta = 0.05$ and $K = 8$, for R_h for Nuclear Norm (NN) regularization, Weighted Nuclear Norm (WNN) regularization, and with linearly-increasing weights (LI) and singular value based (SV) weights. For our proposed method, we also show in red the relative loss improvement compared to the ADMM solution, *i.e.* $100 \times \frac{loss_{ADMM} - loss_{Ours}}{loss_{ADMM}}$.

	Method	NN	WNN	LI	SV
Cathedral (Rank 4)	ADMM	-0.4528	-2.4548	—	—
	Ours	-0.4528 (0%)	-3.8110 (95.6%)	—	—
Door (Rank 4)	ADMM	-1.1946	-1.6247	—	—
	Ours	-1.1946 (0%)	-4.6275 (99.9%)	—	—
École (Rank 4)	ADMM	-0.8747	-1.6869	—	—
	Ours	-0.8747 (0%)	-4.4264 (99.8%)	—	—
Back (Rank 7)	ADMM	-1.3495	-4.8123	-5.2560	-5.2133
	Ours	-1.3536 (0.93%)	-4.8129 (0.14%)	-5.2863 (6.74%)	-5.2471 (7.49%)
Heart (Rank 7)	ADMM	-1.3166	-4.1767	-5.0829	-5.0182
	Ours	-1.3338 (3.88%)	-4.1783 (0.36%)	-5.1153 (7.19%)	-5.0426 (5.46%)
Paper (Rank 7)	ADMM	-1.3298	-5.3847	-5.8484	-5.8509
	Ours	-1.3839 (11.71%)	-5.3910 (1.44%)	-5.8964 (10.46%)	-5.9023 (11.17%)

M_i. Since $(\|B_i\|^2 + \|C_i\|^2)/2$ is non-negative it is clear that selecting z_i as large as possible gives a vector $\mathbf{z}$ that maximizes the right hand side of (18) for all possible (B, C). At iteration t we therefore use

$$r_h^{(t)}(\boldsymbol{\gamma}(B,C)) \approx \sum_{i=1}^{n} w_i^{(t)} \frac{\|B_i\|^2 + \|C_i\|^2}{2} \tag{19}$$

where $w_i^{(t)} = 2(z_i - \eta_i)$ and $z \in \partial g(\boldsymbol{\eta})$ with $z_i = M_i$ when $\eta_i = 0$. Figure 2d shows $r_h(\frac{b^2}{2})$ and the quadratic approximations obtained from the subgradients plotted in 2b. Note that, in contrast to 2b, where we would have to use several subgradients to approximate the functions behaviour around $x = 0$, in 2d it is enough with the one that corresponds to the largest permissible value of z.

2.4. Overview of Algorithm

In this section we give a rough overview of the algorithm we propose (a detailed description is given in the supplementary material). Our algorithm is based on the Variable Projection (VarPro) approach which has been shown to be highly efficient for computer vision problems [18, 21, 19, 17]. Our approximation

$$r_h^{(t)}(\gamma(B,C)) + \|\mathcal{A}(BC)^T - b\|^2, \tag{20}$$

is bi-quadratic meaning that given B we can solve for C in closed form (and vice versa). The optimal $C^*(B)$ as a function of the unknown B can now be inserted back into (20) to give an objective in B alone. VarPro essentially optimizes this new objective locally using a damped Gauss-Newton approach. This is the core routine of our algorithm which consists of the following four main steps:

1. Given $(B^{(t)}, C^{(t)})$ compute the maximal subgradient $\mathbf{z} \in \partial g(\boldsymbol{\gamma}(B^{(t)}, C^{(t)}))$, using the algorithm proposed in [40].

2. Compute the approximation $r_h^{(t)}(\boldsymbol{\gamma}(B,C))$, as in (19).

3. Run one iteration of VarPro on (20) to obtain $(B^{(t+1)}, C^{(t+1)})$.

4. Optional: Compute the SVD $X^{(t+1)} = U\Sigma V^T$, where $X^{(t+1)} = B^{(t+1)}(C^{(t+1)})^T$, and set

$$\begin{aligned} B^{(t+1)} &:= U\sqrt{\Sigma} \\ C^{(t+1)} &:= V\sqrt{\Sigma} \end{aligned} \tag{21}$$

The two first steps give the approximation which is used in VarPro in the third step. The fourth step is optional and is added to help avoid local minima that may occur when $a_i = 0$. In this case the regularizer is typically constant above a certain threshold, as in Figure 2a, and in such cases one may get stuck in suboptimal factorizations. Empirically, we have found that when $a_i \neq 0$ the SVD step can be omitted.

3. Main Theoretical Result

In this section we give the main technical result that makes our algorithmic approach possible.

Theorem 1. *If the sequences $(a_i)_{i=1}^n$ and $(b_i)_{i=1}^n$ are non-negative and non-decreasing then*

$$r_h(\boldsymbol{\sigma}(X)) = \min_{X=BC^T} r_h(\boldsymbol{\gamma}(B,C)) \ . \tag{22}$$

The proof builds on the results of [22] which establishes a similar result but with the function $r_h(\boldsymbol{\sigma}(x))$ replaced by weighted nuclear norm penalties $\mathbf{v}^T\boldsymbol{\sigma}(X)$, where the elements of $\mathbf{v}$ are non-negative and increasing.

3.1. Theoretical Background

Before we proceed to the proof we recall some of the theory from [22]. Here the optimization problem on the right hand side of (22) was studied by writing $X =$

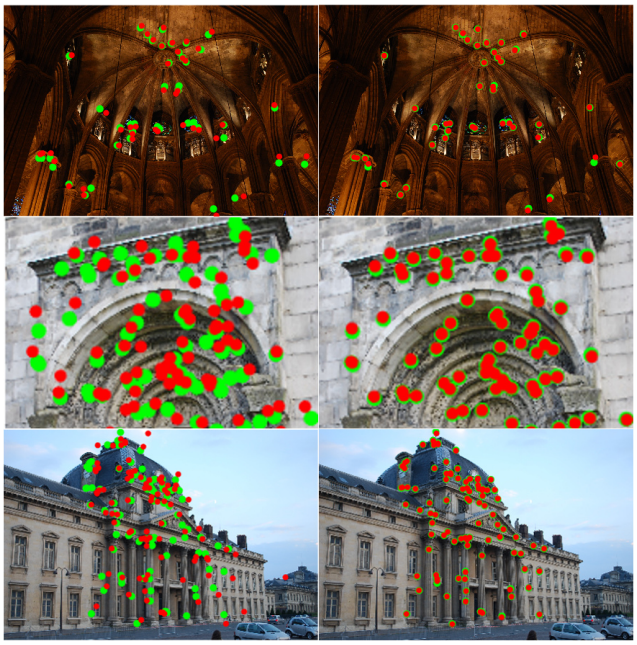


Figure 3: Qualitative results for the Cathedral, Door, and École datasets. In green we show the input image measurements, and in red the reprojected image points. (*Left*): ADMM, and (*Right*): Our method.

BVH^TC^T and varying V and H such that $VH^T = I$. It was shown that the an equivalent formulation is

$$\begin{aligned} \min_{\boldsymbol{\gamma},\, M\in\mathcal{S}} \quad & r_h(\boldsymbol{\gamma}) \\ \text{s.t.} \quad & \boldsymbol{\gamma} = M\boldsymbol{\sigma}, \end{aligned} \tag{23}$$

where

$$\mathcal{S} = \left\{ \frac{1}{2}(V^T \odot V^T + H^T \odot H^T) \mid VH^T = I \right\}. \tag{24}$$

The set $\mathcal{S}$ is difficult to handle since it is non-convex, but it turns out that it is contained in the set of doubly super-stochastic matrices $\mathcal{S}_W$ [3] which is convex. Moreover, the extreme points of $\mathcal{S}_W$ are the permutation matrices which are also contained in $\mathcal{S}$. Therefore optimization of a linear function (weighted nuclear norm) $\mathbf{v}^T\boldsymbol{\gamma}$ over $\mathcal{S}$ is equivalent to optimization over the convex relaxation $\mathcal{S}_W$. Moreover, their results show that if the elements v_i of $\mathbf{v}$ are (strictly) increasing the unique global minimizer is $M = I$.

3.2. Proof of Theorem 1

Our proof relies on the observation that local approximations of $r_h(\boldsymbol{\gamma})$ are equivalent to weighted nuclear norms with with coefficients $v_i = 2(z_i - \gamma_i)$, where $2\mathbf{z} \in \partial g(\boldsymbol{\gamma})$. We will show that these elements are increasing and invoke the results of [22]. For this purpose we need some knowledge about the subgradients $\mathbf{z}$ at $\boldsymbol{\gamma}$. Note that since $\boldsymbol{\gamma}(B, C)$ has non-negative elements we can assume the same for $\mathbf{z}$. Furthermore, because of the permutation invariance of r_h, we also assume that $\boldsymbol{\gamma}$ and $\mathbf{z}$ are non-increasing.

Let $s_i = a_i + \max\{\gamma_i, \sqrt{b_i}\}$. The results of [40] show that if $2\mathbf{z} \in \partial g(\boldsymbol{\gamma})$. Then z_i^* fulfills

$$z_i^* = \begin{cases} s_i & s_i \in [z_{i+1}, z_{i-1}] \\ z_{i-1} & s_i \geq z_{i-1} \\ z_{i+1} & s_i \leq z_{i+1} \end{cases} \tag{25}$$

for all i where $\gamma_i \neq 0$. Note also that the values of these element are the same for all $\mathbf{z}$ with $2\mathbf{z} \in \partial g(\boldsymbol{\gamma})$. For any i where $\gamma_i = 0$, the element z_i can take any value in $[0, s_i]$ as long as the elements of $\mathbf{z}$ are non-increasing. In our case we will use the vector $\mathbf{z}$ that has the largest possible elements and these fulfill (25) for all i. We will now present a fundamental property related to these sequences. A detailed proof is available in the supplementary material.

Lemma 1. *If* $\mathbf{z}^* = \arg\max_{2\mathbf{z}\in\partial g(\boldsymbol{\gamma})} \mathbf{v}^T\mathbf{z}$, *where* $v_i > 0$ *for all* $i = 1, ..., n$ *then the sequence* $(z_i^* - \gamma_i)_{i=1}^n$ *is non-decreasing.*

We now come to the main result.

Proof of Theorem 1. We consider the relaxation of problem (23), namely

$$\begin{aligned} \min_{\boldsymbol{\gamma},\, M\in\mathcal{S}_{\mathcal{W}}} \quad & r_h(\boldsymbol{\gamma}) \\ \text{s.t.} \quad & \boldsymbol{\gamma} = M\boldsymbol{\sigma}. \end{aligned} \tag{26}$$

We will show that minimization of $r_h(\boldsymbol{\gamma})$ over the set $\mathcal{S}_{\mathcal{W}}$ is achieved when $\boldsymbol{\gamma} = \boldsymbol{\sigma}$. Since $\boldsymbol{\sigma} = I\boldsymbol{\sigma}$ and $I \in \mathcal{S}$ this will also be the solution to the original problem.

For simplicity we will first consider the function $\tilde{r}_h(\boldsymbol{\gamma}) = r_h(\boldsymbol{\gamma}) + \langle 2\boldsymbol{w}, \boldsymbol{\gamma}\rangle$, where $\boldsymbol{w} = (\epsilon, 2\epsilon, 3\epsilon, ...)$ for $\epsilon > 0$. The directional derivatives of this function are given by

$$\begin{aligned} (\tilde{r}_h)'_{\boldsymbol{d}}(\boldsymbol{\gamma}) &= g'_{\boldsymbol{d}}(\boldsymbol{\gamma}) - \langle 2\boldsymbol{\gamma}, \boldsymbol{d}\rangle + \langle 2\boldsymbol{w}, \boldsymbol{d}\rangle \\ &= \max_{2z\in\partial g(\gamma)} \langle 2(\mathbf{z} - \boldsymbol{\gamma} + \boldsymbol{w}), \boldsymbol{d}\rangle. \end{aligned} \tag{27}$$

As previously noted, g is convex, which guarantees the existence of the directional derivative. Now consider a point $\boldsymbol{\gamma}^* = M\boldsymbol{\sigma}$, where in $M \in \mathcal{S}_{\mathcal{W}}$. Since r_h is invariant to permutations we may assume that the elements of $\boldsymbol{\gamma}^*$ are non-increasing. We let $2\mathbf{z}^* \in \partial g(\boldsymbol{\gamma}^*)$ with the maximal elements as in Lemma 1. This makes the sequence $z_i^* - \gamma_i^*$ non-decreasing. Therefore the sequence $z_i^* - \gamma_i^* + w_i$ will be strictly increasing. The results of [22] now show that the (unique) minimum of

$$\min_{\boldsymbol{\eta}, M\in\mathcal{S}_{\mathcal{W}}} \quad \langle \mathbf{z}^* - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{\eta}\rangle, \tag{28}$$

$$\text{s.t.} \quad \boldsymbol{\eta} = M\boldsymbol{\sigma}, \tag{29}$$

is given by $\boldsymbol{\eta} = I\boldsymbol{\sigma}$. By selecting $\boldsymbol{d} = \boldsymbol{\sigma} - \boldsymbol{\gamma}^*$ we see that there is a direction such that

$$\langle \mathbf{z}^* - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{d}\rangle = \langle \mathbf{z}^* - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{\sigma}\rangle - \langle \mathbf{z}^* - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{\gamma}^*\rangle < 0. \tag{30}$$

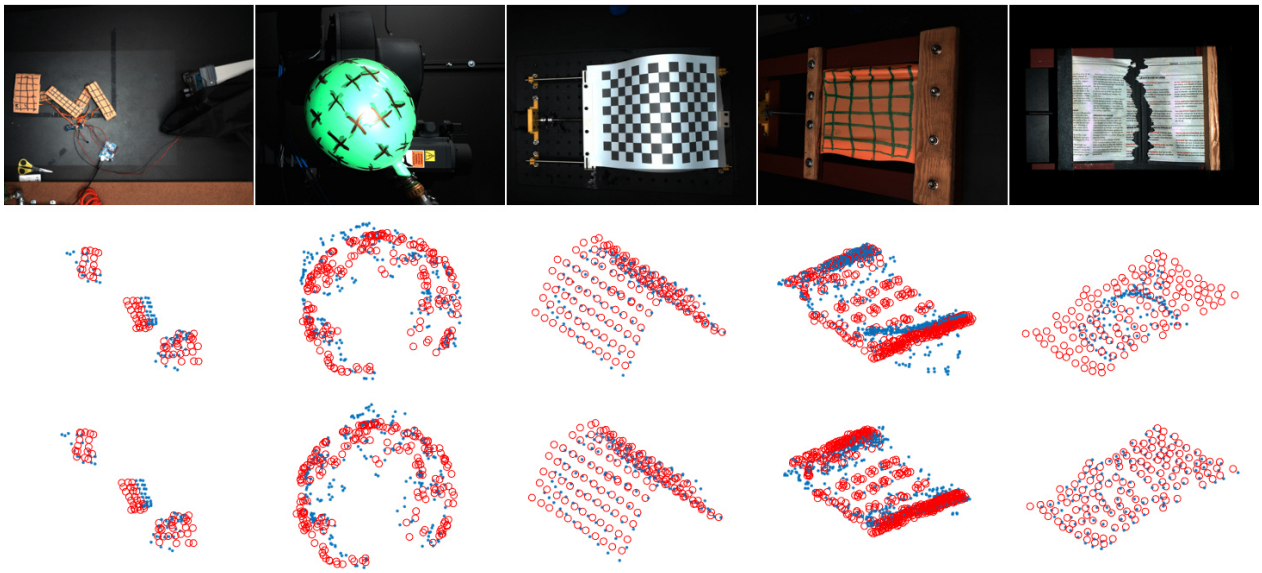


Figure 4: Qualitative results on the (left to right) Articulated, Balloon, Paper, Stretch, and Tearing datasets of the NRSfM Challenge. In red we show the provided ground-truth structure for one of the frames of the sequences. In blue we show the obtained 3D structure for the same frame using ADMM (top) and the proposed method (bottom).

Furthermore, for any other choice $\mathbf{z} \in \partial g(\boldsymbol{\gamma}^*)$ we have $\langle \mathbf{z}^*, \boldsymbol{\sigma} \rangle \geq \langle \mathbf{z}, \boldsymbol{\sigma} \rangle$ since the elements of $\boldsymbol{\sigma}$ are non-negative and $\langle \mathbf{z}^*, \boldsymbol{\gamma}^* \rangle = \langle \mathbf{z}, \boldsymbol{\gamma}^* \rangle$ since $z_i^* = z_i$ when $\gamma_i \neq 0$. Therefore

$$\langle \mathbf{z}^* - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{d} \rangle \geq \langle \mathbf{z} - \boldsymbol{\gamma}^* + \boldsymbol{w}, \boldsymbol{d} \rangle, \tag{31}$$

which shows that $\mathbf{z}^*$ is the maximizer in (27).

We can thus conclude that, as long as $\boldsymbol{\gamma}^* \neq \boldsymbol{\sigma}$, there is a direction (feasible in $\mathcal{S}_{\mathcal{W}}$) with strictly negative directional derivative. Hence $\boldsymbol{\sigma}$ is the minimizer of $\tilde{r}_h(\boldsymbol{\gamma})$ over $\mathcal{S}_{\mathcal{W}}$. Since $I \in \mathcal{S}$ it also optimizes $\tilde{r}_h(\boldsymbol{\gamma})$ over $\mathcal{S}$ for any $\epsilon > 0$.

We now let $\epsilon \to 0$ to show that $\boldsymbol{\sigma}$ is a (not necessarily unique) minimizer of (26). To see this we note that since $\langle \boldsymbol{w}, \boldsymbol{\gamma} \rangle > 0$ for all feasible $\boldsymbol{\gamma}$ we have

$$r_h(\boldsymbol{\sigma}) < r_h(\boldsymbol{\sigma}) + 2\langle \boldsymbol{w}, \boldsymbol{\sigma} \rangle \leq r_h(\boldsymbol{\gamma}) + 2\langle \boldsymbol{w}, \boldsymbol{\gamma} \rangle. \tag{32}$$

Taking the (pointwise) limit of the right hand side shows that $r_h(\boldsymbol{\sigma}) \leq r_h(\boldsymbol{\gamma})$. □

4. Experiments

4.1. Matrix Recovery

In this section, the convergence and accuracy of the proposed method, using bilinear formulation, are compared to the ADMM approach [40]. For the data term, we use the pseudo Object Space Error (pOSE) defined in Hong *et al*. [21], which fuse the Object Space Error (OSE)

$$\ell_{\text{OSE}} \coloneqq \sum_{(i,j)\in\Omega} \|\mathrm{P}_{i,1:2}\tilde{\mathbf{x}}_j - (\mathbf{p}_{i,3}^T\tilde{\mathbf{x}}_j)\mathbf{m}_{i,j}\|_2^2, \tag{33}$$

and the error of an affine camera model

$$\ell_{\text{Affine}} \coloneqq \sum_{(i,j)\in\Omega} \|\mathrm{P}_{i,1:2}\tilde{\mathbf{x}}_j - \mathbf{m}_{i,j}\|_2^2, \tag{34}$$

where $\mathrm{P}_{i,1:2}$ and $\mathbf{p}_{i,3}$ are, respectively, the first two and the third rows of the camera matrix P_i, with $i = 1, \ldots, F$. Furthermore, $\tilde{\mathbf{x}}_j$ is a 3D point in homogeneous coordinates, with $j = 1, \ldots, P$, and $\mathbf{m}_{i,j}$ is the 2D observation of the j:th point on the i:th camera. The set of observable data is denoted Ω. The two terms are weighted by $\eta \in [0, 1]$, resulting in the loss

$$\ell_{\text{pOSE}} \coloneqq (1-\eta)\ell_{\text{OSE}} + \eta\ell_{\text{Affine}}. \tag{35}$$

In Iglesias *et al*. [22], the possibility of extending this framework to non-rigid structure from motion is described by replacing $\mathrm{P}_i\tilde{\mathbf{x}}_j$ by a linear combination of K shape basis, *i.e*. $\Pi_i\hat{\mathrm{S}}_j$. These factors are structured as $\Pi_i = \begin{bmatrix} c_{i,1}R_i & \ldots & c_{i,K}R_i & t_i \end{bmatrix} \in \mathbb{R}^{3\times(3K+1)}$ and $\hat{\mathrm{S}}_j = \begin{bmatrix} S_{1,j}^T & \ldots & S_{K,j}^T & 1 \end{bmatrix}^T \in \mathbb{R}^{3K+1}$. The pOSE objective can now be written in a more compact form as

$$\ell_{\text{pOSE}}(X) = \|\mathcal{A}(X) - b\|^2, \tag{36}$$

where $X = \Pi\hat{\mathrm{S}}$, and Π and $\hat{\mathrm{S}}$ are the vertical and horizontal concatenations of Π_i and $\hat{\mathrm{S}}_j$, respectively. For the ADMM formulation we consider the optimization problem

$$\min_X \quad \mathcal{R}_h(X) + \ell_{\text{pOSE}}(X), \tag{37}$$

while, for our proposed method, we have the equivalent problem

$$\min_{B,C} \quad \tilde{\mathcal{R}}_h(B, C) + \ell_{\text{pOSE}}(BC^T). \tag{38}$$

The datasets chosen for this experiment are Cathedral, École, and Door [32] (rigid scenes), Back [35], Heart [37], and Paper [42] (deformable objects). We use four different sets of sequences (a_i) and (b_i): (i) nuclear norm (NN), with $a_i = a_{NN}$, and $b_i = 0$, (ii) weighted nuclear norm (WNN), with $a_i = {}^{a_{WNN}}/_{(\sigma_i(X_0)+\delta)}$, and $b_i = 0$, where $\delta > 0$, (iii) linearly increasing (LI) weights with $a_i = b_i = 0, 0 \geq i \geq 4$, and $a_i = 10b_i = a_{LI}(i-4), i \geq 5$, and (iv) singular value (SV) based sequences $a_i = 2b_i = {}^{a_{SV}}/_{(\sigma_i(X_0)+\delta)}$. The singular values $\boldsymbol{\sigma}(X_0)$ are obtained from the solution X_0 of $\|\mathcal{A}(X) - b\|^2$ (with no regularization). We set $k = 8$ and the parameters for the sequences are chosen such that the methods converge to rank 4 and 7 solutions for the rigid and deformable datasets, respectively. The values of the parameters used to define the sequences for each of the datasets are shown in the supplementary material. For increased stability regarding parameter choice across datasets, we normalize the measurement matrix M to have unit Frobenius norm.

In Table 1 we show the losses obtained by both methods, where the advantages of using our bilinear formulation become clear. In Figure 3 we also show qualitative results of the estimated reprojections on the Cathedral, Door, and École datasets. A convergence plot for the Door dataset is shown in Figure 5, and for the remaining datasets we refer to the supplementary material. The results show the improvement in accuracy that can be obtained by using the bilinear formulation. The difference in terms of loss after convergence, and how it affects the reprojection errors, is more noticeable in the rigid datasets. However, in the next section we show that even a small improvement in the reprojection errors can lead to significantly better 3D reconstructions.

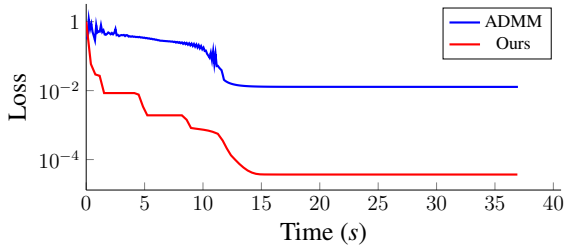


Figure 5: Convergence history for the Door dataset.

4.2. Application to Non-Rigid SfM

We apply our method to the perspective datasets in the Non-Rigid Structure from Motion Challenge [23]. Each of the five datasets—Articulated, Balloon, Paper, Stretch and Tearing—have six different sequences, consisting of six different camera paths—circle, flyby, line, semi-circle, tricky and zigzag. To deal with perspective projections, we again use the pseudo Object Space Error $\ell_{\text{pOSE}}(X)$ described in Section 4.1, where X is now parameterized as $R\mathcal{U}(CB^\sharp) + t$, with t being the vector of translations. Similarly to Iglesias *et al.* [22], we assume that the rotations are known, and that the image measurements are calibrated. The rotation matrices are recovered from 2D observations for the orthogonal camera model, while the intrinsic camera matrix is estimated using the provided ground-truth 3D structure for one of the frames of the sequences. The optimization problem can be written as

$$\min_{X,t} \quad \mathcal{R}_h(X^\sharp) + \ell_{\text{pOSE}}(RX + t), \tag{39}$$

and with our bilinear formulation we solve

$$\min_{C,B^\sharp,t} \quad \tilde{\mathcal{R}}_h(C, B^\sharp) + \ell_{\text{pOSE}}(R\mathcal{U}(CB^\sharp) + t). \tag{40}$$

In our comparison, we employ the same initialization heuristic for the weights w_i on the singular values as in [24], namely

$$w_i = \frac{\xi}{\sigma_i(X_0^\sharp) + \delta}, \tag{41}$$

where $\delta > 0$ is added to avoid division by zero, and $\xi > 0$. The matrix $X_0^\sharp = R^+M$, where R^+ is the pseudo-inverse of R, which is a common initialization scheme for NRSfM [11, 39, 24].

We use $\eta = 0.05$ and $K = 3$, while for the sequences, we choose $a_i = w_i$ and $b_i = w_i/2$, where w_i is defined in (41) with $\xi = 5 \times 10^{-3}$ and $\delta = 10^{-8}$. We refine the solution obtained with ADMM formulation (after convergence) by performing 100 iterations of our method. The results are summarized in Table 2, with the average log-losses and 3D reconstruction errors, respectively, over the 6 different sequences of each dataset. In Figure 4 we show qualitative results for one the sequences of each dataset, as well as examples of the convergence plots of the two methods. Our method was always able to achieve a better loss than the ADMM solution, and in 4 out of the 5 datasets that improvement in loss led to a more accurate 3D reconstruction.

Table 2: Average log-loss and reconstruction errors (mm) on each dataset over the 6 camera paths relatively to the provided ground-truth structure.

		Artic.	Balloon	Paper	Stretch	Tearing
Log-loss	ADMM	-2.221	-2.529	-2.338	-2.395	-1.471
	Ours	**-2.415**	**-2.657**	**-2.560**	**-2.622**	**-2.053**
Rec. error	ADMM	**14.55**	9.29	6.95	7.83	29.90
	Ours	16.10	**8.29**	**6.70**	**7.66**	**11.26**

5. Conclusions

In this paper we have provided a bilinear optimization framework compatible with a wide range of penalty functions. Furthermore, we have shown that the proposed regularizer is equivalent with the linear counterpart, making the transition from a splitting scheme based methodology to our differentiable bilinear framework easier.

References

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

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

[3] Rajendra Bhatia and Tanvi Jain. On symplectic eigenvalues of positive definite matrices. *Journal of Mathematical Physics*, 56(11):112201, 2015. 6

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011. 2

[5] 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

[6] 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

[7] 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. 1, 2, 3

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

[9] Lu Canyi, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Generalized nonconvex nonsmooth low-rank minimization. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2

[10] Marcus Carlsson. On convex envelopes and regularization of non-convex functionals without moving global minima. *Journal of Optimization Theory and Applications, to appear*, 2019. 2, 3, 4

[11] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122, 2014. 1, 2, 8

[12] Anders Eriksson, Trung Thanh Pham, Tat-Jun Chin, and Ian Reid. The k-support norm and convex envelopes of cardinality and rank. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[13] Christian Grussler and Pontus Giselsson. Low-rank inducing norms with optimality interpretations. *CoRR*, abs/1612.03186, 2016. 2

[14] Christian Grussler, Anders Rantzer, and Pontus Giselsson. Low-rank optimization with convex constraints. *CoRR*, abs/1606.01793, 2016. 2

[15] Benjamin D. Haeffele and René Vidal. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *CoRR*, abs/1708.07850, 2017. 1, 3

[16] J. He, L. Balzano, and A. Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1568–1575, June 2012. 2

[17] Je Hyeong Hong and Andrew Fitzgibbon. Secrets of matrix factorization: Approximations, numerics, manifold optimization and random restarts. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5

[18] Je Hyeong Hong, Christopher Zach, and Andrew Fitzgibbon. Revisiting the variable projection method for separable nonlinear least squares problems. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 5

[19] Je Hyeong Hong, Christopher Zach, Andrew Fitzgibbon, and Roberto Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 5

[20] 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. 1, 2

[21] Je Hyeong Hong and Christopher 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, 2, 5, 7

[22] José Pedro Iglesias, Carl Olsson, and Marcus Valtonen Örnhag. Accurate optimization of weighted nuclear norm for non-rigid structure from motion. In *Proceedings of the European Computer Vision Conference (ECCV)*, 2020. 3, 5, 6, 7, 8

[23] Sebastian Hoppe Nesgaard Jensen, Alessio Del Bue, Mads Emil Brix Doest, and Henrik Aanæs. A benchmark and evaluation of non-rigid structure from motion, 2018. 8

[24] Suryansh Kumar. A simple prior-free method for non-rigid structure-from-motion factorization : Revisited. *CoRR*, abs/1902.10274, 2019. 1, 8

[25] Suryansh Kumar. Non-rigid structure from motion: Prior-free factorization method revisited. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 51–60. IEEE, 2020. 1, 2

[26] Hanjiang Lai, Yan Pan, Canyi Lu, Yong Tang, and Shuicheng Yan. Efficient k-support matrix pursuit. In *European Conference on Computer Vision 2014*, volume 8690, 2014. 2

[27] Canyi Lu, Changbo Zhu, Chunyan Xu, Shuicheng Yan, and Zhouchen Lin. Generalized singular value thresholding. In *AAAI*, 2015. 1, 2

[28] Andrew M McDonald, Massimiliano Pontil, and Dimitris Stamos. Spectral k-support norm regularization. In *Advances in Neural Information Processing Systems*. 2014. 2

[29] 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. 1, 2

[30] Carl Olsson, Marcus Carlsson, Fredrik Andersson, and Viktor Larsson. Non-convex rank/sparsity regularization and local minima. *Proceedings of the International Conference on Computer Vision*, 2017. 4

[31] C. Olsson, M. Carlsson, and E. Bylow. A non-convex relaxation for fixed-rank approximation. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1809–1817, 2017. 4

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

[33] Marcus Valtonen Örnhag, Carl Olsson, and Anders Heyden. Differentiable fixed-rank regularisationusing bilinear parameterisation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019. 3

[34] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, Aug. 2010. 2, 4

[35] Chris Russell, Joao Fayad, and Lourdes Agapito. Energy based multiple model fitting for non-rigid structure from motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009 – 3016, 07 2011. 8

[36] 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. 3

[37] Danail Stoyanov, George P. Mylonas, Fani Deligianni, Ara Darzi, and Guang Zhong Yang. Soft-tissue motion tracking and structure estimation for robotic assisted mis procedures. In James S. Duncan and Guido Gerig, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, pages 139–146, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. 8

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

[39] J. Valmadre, S. Sridharan, S. Denman, C. Fookes, and S. Lucey. Closed-form solutions for low-rank non-rigid reconstruction. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6, Nov 2015. 8

[40] Marcus Valtonen Örnhag and Carl Olsson. A unified optimization framework for low-rank inducing penalties. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 4, 5, 6, 7

[41] Marcus Valtonen Örnhag, Carl Olsson, and Anders Heyden. Bilinear parameterization for differentiable rank-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 3

[42] Aydin Varol, Mathieu Salzmann, Engin Tola, and Pascal Fua. Template-free monocular reconstruction of deformable surfaces. In *International Conference on Computer Vision (ICCV)*, pages 1811 – 1818, 11 2009. 8

[43] Naiyan Wang, Tiansheng Yao, Jingdong Wang, and Dit-Yan Yeung. A probabilistic approach to robust matrix factorization. In *European Conferencec on Computer Vision (ECCV)*, 2012. 1

[44] Chen Xu, Zhouchen Lin, and Hongbin Zha. A unified convex surrogate for the schatten-p norm. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2017. 3

[45] Noam Yair and Tomer Michaeli. Multi-scale weighted nuclear norm image restoration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1