Differentiable Fixed-Rank Regularisation using Bilinear Parameterisation

Marcus Valtonen Örnhag[1]
marcus.valtonen_ornhag@math.lth.se
Carl Olsson[1,2]
caols@chalmers.se
Anders Heyden[1]
heyden@maths.lth.se

[1] Centre for Mathematical Sciences
Lund University
Lund, Sweden

[2] Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden

Abstract

Low rank structures are present in many applications of computer vision and machine learning. A popular approach consists of explicitly parameterising the set or matrices with sought rank, leading to a bilinear factorisation, reducing the problem to find the bilinear factors. While such an approach can be efficiently implemented using second-order methods, such as Levenberg–Marquardt (LM) or Variable Projection (VarPro), it suffers from the presence of local minima, which makes theoretical optimality guarantees hard to derive.

Another approach is to penalise non-zero singular values to enforce a low-rank structure. In certain cases, global optimality guarantees are known; however, such methods often lead to non-differentiable (and even discontinuous) objectives, for which it is necessary to use subgradient methods and splitting schemes. If the objective is complex, such as in structure from motion, the convergence rates for such methods can be very slow.

In this paper we show how optimality guarantees can be lifted to methods that employ bilinear parameterisation when the sought rank is known. Using this approach the best of two worlds are combined: optimality guarantees and superior convergence speeds. We compare the proposed method to state-of-the-art solvers for prior-free non-rigid structure from motion.

1 Introduction

The singular value decomposition (SVD) has long been the main tool for enforcing rank constraint. It is well-known that when all elements are measured the optimal solution is obtained by finding the SVD of a matrix, thresholding the first k singular values, and recombining the remaining entries to obtain the optimal rank k approximation (measured in the Frobenius norm). In computer vision and machine learning applications, however, missing data patterns emerge, often in structured ways, and such simple methods are not directly applicable.

In this paper we will consider low rank approximation problems with structured missing data problems, which *e.g.* emerge in structure from motion problems. More specifically, we are interested in solving

$$\min_{\operatorname{rank}(X) \leq r_0} \|\mathcal{A}(X) - b\|^2. \tag{1}$$

where $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ is a linear operator, $b \in \mathbb{R}^p$ and $\| \cdot \|$ is the standard Euclidean norm. We may, equivalently, consider the unconstrained problem formulation

$$\min_X \mathbb{I}(\operatorname{rank}(X) \leq r_0) + \|\mathcal{A}(X) - b\|^2. \tag{2}$$

where $\mathbb{I}(\operatorname{rank}(X) \leq r_0)$ is the indicator function attaining the value 0 if $\operatorname{rank}(X) \leq r_0$ and ∞ otherwise. Due to being discontinuous, the indicator function is not suitable for numerics. The standard approach has been to relax it with a convex alternative, such as the nuclear norm $\|X\|_* = \sum_{i=1}^n \sigma_i(X)$, where $\sigma_i(X)$ is the i:th singular value of X. This relaxation has the theoretical benefit of being the convex envelope of the rank function over the set $\{X : \sigma_1(X) \leq 1\}$, see *e.g.* [10], which further has led to generalisations with performance guarantees [5, 6, 22, 24]. Due to the fact that the singular values are penalised equally hard, regardless of size, the nuclear norm has a shrinking bias [20]. For computer vision problems this is often an undesirable effect, which has been shown on structure-from-motion problems [8, 19]. Instead, non-convex relaxations that penalise smaller singular values harder than larger ones have been shown to improve performance [14, 16, 17, 21]. One of the downsides of using these types of regularisation terms is that they often result in non-differentiable objectives, which makes it necessary to use splitting schemes or subgradient based methods. This in turn may affect the convergence rates, making it infeasible for certain types of problems.

Since the sought rank is known, we may instead consider optimising over a bilinear factorisation $X = BC^T$, where $B \in \mathbb{R}^{m \times r}$ and $C \in \mathbb{R}^{n \times r}$, which explicitly parameterises the family of matrices of size $m \times n$ with rank at most r. While methods utilising a bilinear parameterisation can be extended to second order methods, with fast convergence rates in the neighbourhood of local minima, they instead suffer from the presence of local minima. Therefore, attempts to unify regularisation terms with bilinear factorisation has gained some attention in recent years [2, 4, 20, 25, 27], with the hope of combining the best from both worlds: theoretical performance guarantees and fast convergence.

2 Related Work

Cabral *et al.* [4] considered the variational formulation of the nuclear norm

$$\|X\|_* = \min_{BC^T = X} \frac{1}{2}(\|B\|_F^2 + \|C\|_F^2), \tag{3}$$

see [24], and unified the use of a regularised objective and bilinear factorisation. They are able to prove global optimality in cases where the obtained solution has lower rank than the number of columns. Bach [2] extended this property to regularisers on the form $\|X\|_{s,t} = \min_{BC^T = X} \frac{1}{2} \sum_{i=1}^k (\|B_i\|_s^2 + \|C_i\|_t^2)$, where $\| \cdot \|_p$ is the ℓ_p-norm. These methods, however, rely on convexity, thus implicitly suffers from shrinking bias. Shang *et al.* [25] went beyond convexity and studied the Schatten semi-norms $\|X\|_q = \sqrt[q]{\sum_{i=1}^N \sigma_i(X)^q}$, for $q = 1/2$ and $q = 2/3$ (which was generalised in [27]). The shrinking bias was drastically reduced, but the proposed method does not benefit from the convergence rates of second-order methods, as the non-convexity is treated using a splitting scheme.

Recently, Valtonen Örnhag *et al.* [20] studied a family of regularisers on the form $\mathcal{R}(X) = \sum_{i=1}^{k} f(\sigma_i(X))$, where f is a robust penalty function, assumed to be concave and nondecreasing on $[0,\infty)$ with $f(0) = 0$. They showed that objectives incorporating such regularisers, which are non-differentiable by nature, can be reformulated into differentiable objectives using bilinear factorisation. Furthermore, they can be optimised using second-order methods such as Levenberg–Marquardt (LM) and Variable Projection (VarPro). Theoretical optimality guarantees for the choice $f(x) = f_\mu(x) := \mu - \max(\sqrt{\mu} - x, 0)$ which is equivalent to the the convex envelope of the "soft rank" objective $\mu \operatorname{rank}(X) + \|X - M\|_F^2$ were studied. Previously known results, including global optimality guarantees, were transferred to the bilinear setting.

In this paper we will focus on overparameterising bilinear formulations for the "hard rank" objective (2). In [15] a special case was studied

$$\min_X \mathbb{I}(\operatorname{rank}(X) \leq r_0) + \|X - M\|_F^2, \tag{4}$$

and it was shown that the convex envelope of the objective function is given by

$$\mathcal{R}_{r_0}(X) + \|X - M\|_F^2, \tag{5}$$

where

$$\mathcal{R}_{r_0}(X) = \max_Z \sum_{i=r_0+1}^{n} \sigma_i(Z)^2 - \|X - Z\|_F^2, \tag{6}$$

Since (5) is the convex envelope of (4) the global minimisers are attained simultaneously. Furthermore, (5) is continuous and convex, which makes it tractable for practical problems. Inspired by the approach used in Valtonen Örnhag *et al.* [20] we show that it is possible to combine theoretical optimality guarantees and second order methods for the hard rank objective as well. Our contributions are:

- A novel method for regularising fixed-rank problems,
- Optimality guarantees for a wide range of problems,
- Comparison to state-of-the-art methods on Non-Rigid Structure from Motion.

3 Differentiable Regularisers

In this section we will derive an alternative regulariser to (6) with certain desired properties suitable in a framework utilising bilinear parameterisation. One such property is differentiability.

3.1 Bilinear Parameterisation and Pseudo-Singular Values

Let $X \in \mathbb{R}^{m\times n}$, with $\operatorname{rank}(X) = r_0$. Then there exists a decomposition $B \in \mathbb{R}^{n\times k}$ and $C \in \mathbb{R}^{m\times k}$, with $r_0 \leq k$, such that $X = BC^T$. Furthermore, define the *pseudo-singular values*

$$\gamma_i(B,C) := \frac{\|B_{[i]}\|^2 + \|C_{[i]}\|^2}{2}, \tag{7}$$

where the square brackets indicate the columns of B and C such that the pseudo-singular values are sorted in descending order $\gamma_1(B,C) \geq \cdots \geq \gamma_k(B,C) \geq 0$. Note that if $X = U\Sigma V^T$

is a SVD of X, then the re-factorisation $B = U\sqrt{\Sigma}$ and $C = V\sqrt{\Sigma}$, such that $X = BC^T$, has the properties that the singular values and the pseudo-singular values coincide $\gamma_i(B,C) = \sigma_i(X)$, for all $i = 1,\dots,k$, if we use the convention that $\sigma_i(X) = 0$ for $i > r_0$.

3.2 A Bilinear Fixed-Rank Regulariser

Let $\sigma(X)$ denote the singular value vector of X, and note that the first term of (6) is unitarily invariant, whereas the second term can be expressed as $-\|X-Z\|_F^2 = 2\langle X,Z\rangle_F - \|X\|_F^2 - \|Z\|_F^2$. Recall that, by von Neumann's trace theorem, $|\langle X,Z\rangle_F| \leq \langle \sigma(X), \sigma(Z)\rangle$, with equality when X and Z are simultaneously unitarily diagonalisable. This reduces the problem to maximising over the singular values alone,

$$\mathcal{R}_{r_0}(X) = \max_{\sigma(Z)} \left(\sum_{i=r_0+1}^{n} \sigma_i^2(Z) - \sum_{i=1}^{n} (\sigma_i(Z) - \sigma_i(X))^2 \right). \tag{8}$$

A bilinear formulation equivalent to (8) can now be created by replacing the singular values by the pseudo-singular values (7), which yields

$$\tilde{\mathcal{R}}_{r_0}(B,C) := \max_{z\in\mathcal{Z}} \left(\sum_{i=r_0+1}^{n} z_i^2 - \sum_{i=1}^{n} (z_i - \gamma_i(B,C))^2 \right), \tag{9}$$

where $\mathcal{Z} = \{z : z_1 \geq \cdots \geq z_n \geq 0\}$. Clearly, $\mathcal{R}_{r_0}(X) = 0$ if and only if $\operatorname{rank}(X) \leq r_0$; hence, the proposed regulariser will penalise solutions admitting more than r_0 non-zero columns in the bilinear factorisation. By recomputing the bilinear factors such that $B = U\sqrt{\Sigma}$ and $C = V\sqrt{\Sigma}$, where $X = U\Sigma V^T$ is a SVD, yields $\tilde{\mathcal{R}}_{r_0}(B,C) = \mathcal{R}_{r_0}(BC^T)$, thus, implicitly, enforces the rank constraint.

3.3 Differentiability

We next show that the cost function, including the bilinear regulariser (9), is differentiable.

Theorem 1. *The function $\mathcal{F} : \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} \to \mathbb{R}$, defined as*

$$\mathcal{F}(B,C) = \tilde{\mathcal{R}}_{r_0}(B,C) + \|\mathcal{A}(BC^T) - b\|^2, \tag{10}$$

is differentiable w.r.t. B and C.

Proof sketch. For a complete proof, see the supplementary material. Decompose $\mathcal{F}(B,C) = G(B,C) + H(B,C)$ where $G(B,C) = \tilde{\mathcal{R}}_{r_0}(B,C) + \sum_{i=1}^{k} \gamma_i^2(B,C)$ and $H(B,C) = -\sum_{i=1}^{k} \gamma_i^2(B,C) + \|\mathcal{A}(BC^T) - b\|^2$. Clearly, H is differentiable, and we will show that G is convex and differentiable. Let $\gamma : \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} \to \mathbb{R}^k$ denote the function that takes the bilinear factors and returns the pseudo-singular values,

$$\gamma(B,C) = \frac{1}{2}\left(\|B_{[1]}\|^2 + \|C_{[1]}\|^2, \dots, \|B_{[k]}\|^2 + \|C_{[k]}\|^2\right). \tag{11}$$

and

$$\phi(\gamma) = \max_{z_1\geq z_2\geq \dots \geq z_k \geq 0} L(\gamma,z)^T, \tag{12}$$

where $L(\gamma,z) = -\sum_{i=1}^{r_0} z_i^2 + 2\sum_{i=1}^{k} z_i\gamma_i$. Note that $G(B,C) = \phi(\gamma(B,C))$ and by the chain rule $\partial_B G(B,C) = \nabla_B \gamma(B,C)\partial_\gamma \phi(\gamma(B,C))$. Here $\nabla_B \gamma(B,C) = \operatorname{blkdiag}(B_{[1]}, B_{[2]}, \dots, B_{[k]})$, $\partial_\gamma \phi =$

$\left(\partial_{\gamma_{[1]}}\phi(\gamma), \partial_{\gamma_{[2]}}\phi(\gamma), \ldots, \partial_{\gamma_{[k]}}\phi(\gamma)\right)^T$ and therefore $\partial_B G(B,C)$ is an $mk \times 1$ vector containing derivatives with respect to the elements of B. Using a similar argument as in [15], the subdifferential can be written as

$$\partial\phi(\gamma) = 2 \underset{z_1 \geq \cdots \geq z_k \geq 0}{\arg\max} \{L(\gamma, z)\}. \tag{13}$$

Analogously to [15], the optimising vector z is given by

$$z_i \in \begin{cases} \{\max(\gamma_i, s)\}, & i \leq r_0 \\ \{s\}, & i \geq r_0,\ \gamma_i \neq 0 \\ [0, s], & i > r_0,\ \gamma_i = 0 \end{cases} \tag{14}$$

for some $s \geq \gamma_{r_0}$. For G to be differentiable with respect to B it is sufficient that $\partial_B G(B,C)$ contains a single element. To see that this is true one notes that only the elements of the subgradient $\partial\phi(\gamma)$, for which $\gamma_i = 0$ and $i > r_0$, can have non-singleton sets; however, $\gamma_i = 0$ implies that both $B_{[i]}$ and $C_{[i]}$ are zero vectors. Therefore all elements $\partial_{\gamma_{[i]}}\phi(\gamma)$ that can take multiple values vanish in the multiplication $\nabla_B\gamma(B,C)\partial_\gamma\phi(\gamma(B,C))$. In conclusion, G is differentiable with respect to B, thus also $\mathcal{F}$. An identical argument now shows that the same is true for derivatives with respect to C. □

4 Optimality Conditions

In this section, we show that previously known results concerning the regulariser (6) can be transferred to the bilinear setting. This is done by first establishing a crucial relation between $\mathcal{R}_{r_0}(X)$ and the proposed regulariser $\tilde{\mathcal{R}}_{r_0}(B,C)$. This is a generalisation of [20], and we use the same strategy, namely, to overparameterise the bilinear factorisation such that B and C have $2k$ columns. This, in turn, allows us to parameterise line segments between points of at most rank k, in addition to applying convexity properties.

It should be noted that overparameterisation introduces additional stationary points. Considering the data term $\|\mathcal{A}(BC^T) - b\|^2$, it is clear that the gradients w.r.t. the bilinear factors vanish at $(B,C) = (0,0)$. This is not the case for gradients w.r.t. $X = BC^T$ which are nonzero, in general. Nevertheless, it is possible to relate the local minima of the problem formulation (5) and the directional derivatives between low rank factorisations of the proposed bilinear formulation, which is done in Theorem 2.

Theorem 2. *Assume that $(\bar{B},\bar{C}) \in \mathbb{R}^{m\times 2k} \times \mathbb{R}^{n\times 2k}$ is a local minimizer of (10), where $\bar{B} = U\sqrt{\Sigma}$ and $\bar{C} = V\sqrt{\Sigma}$, and $\bar{X} = U\Sigma V^T$, with $r_0 < k$ non-zero columns and let $\mathcal{N}(X) = \mathcal{R}_{r_0}(X) + \|\mathcal{A}(X) - b\|^2$. If $\mathcal{R}_{r_0}(\bar{X}) = \tilde{\mathcal{R}}_{r_0}(\bar{B},\bar{C})$ then the directional derivatives $\mathcal{N}'_{\Delta X}(\bar{X})$, where $\Delta X = \tilde{X} - \bar{X}$, $\operatorname{rank}(\tilde{X}) \leq k$ are non-negative.*

The assumptions that the local minimizers must be on the form $\bar{B} = U\sqrt{\Sigma}$ and $\bar{C} = V\sqrt{\Sigma}$, respectively, is not restricting the applicability, since there can be local minimizers for which $\tilde{\mathcal{R}}_{r_0}(\bar{B},\bar{C}) \neq \mathcal{R}_{r_0}(\bar{B}\bar{C}^T)$. In order to avoid such situations, we recompute the SVD each iteration, according to $\bar{B} = U\sqrt{\Sigma}$ and $\bar{C} = V\sqrt{\Sigma}$, assuming $U\Sigma V^T$ is an SVD of $\bar{B}\bar{C}^T$. By doing so, we enforce $\tilde{\mathcal{R}}_{r_0}(\bar{B},\bar{C}) = \mathcal{R}_{r_0}(\bar{B}\bar{C}^T)$.

Using the result from Theorem 2, we can apply the non-negativity argument of the directional derivatives, to problems obeying the *restricted isometry property* (RIP) [24]

$$(1 - \delta_{2k})\|X\|_F^2 \leq \|\mathcal{A}(X)\|^2 \leq (1 + \delta_{2k})\|X\|_F^2, \tag{15}$$

for some $0 < \delta_{2k} < 1$. The theory of local minimizers in the bilinear factors of $\|\mathcal{A}(BC^T) - b\|^2$ under the RIP constraint has recently been developed in [3, 11, 23]. This is a well-studied class of problems, for which it can be shown that no spurious local minima are introduced using the bilinear factorisation [23] (under standard regulatory assumptions on $\mathcal{A}$). The fundamental idea is to bound the distance between the global minimum and the local minima, which can be made small in terms of the residual error; however, this approach does not, in general, guarantee uniqueness of local minima. By lifting the results of [18] to the bilinear setting, we can in fact prove uniqueness under the RIP constraint. Theorem 3 shows how this can be incorporated to ensure global minimality.

Theorem 3. *Assume that $(\bar{B},\bar{C})$ is a local minimizer of (10) fulfilling the assumptions of Theorem 2. If the singular values of $Z = (I - \mathcal{A}^*\mathcal{A})\bar{B}\bar{C}^T + \mathcal{A}^*b$, where $\mathcal{A}^*$ is the adjoint operator of $\mathcal{A}$, fulfil $\sigma_{r_0+1}(Z) < (1-2\delta_{2r_0})\sigma_{r_0}(Z)$, then*

$$\bar{B}\bar{C}^T \in \underset{\mathrm{rank}(X)\leq r_0}{\arg\min} \; \|\mathcal{A}(X) - b\|^2. \tag{16}$$

In most practical situations, the separation of the singular values of Z is easily fulfilled. Consider, *e.g.* the case of exact data, when $b = \mathcal{A}(X_0)$, where $\mathrm{rank}(X_0) = r_0$. Then $Z = X_0$, and the condition is transferred to the properties of the local (global) minima X_0, *i.e.* $\sigma_{r_0+1}(X_0) < (1-2\delta_{2r_0})\sigma_{r_0}(X_0)$. The proof is given in the supplementary material.

5 Implementation

The method we propose is based on VarPro [12], which allows us to work with the low-rank factors B and C directly. Assuming a separable non-linear least squares problem, in the components $b = \mathrm{vec}(B)$ and $c = \mathrm{vec}(C^T)$, VarPro reduces the problem by solving a nonlinear problem in b only. This is done by marginalising c, prior to optimising over b. For a thorough overview of VarPro for computer vision applications, see [13].

Many computer vision and machine learning problems are large, but highly structured. Failure to retain any previous structure by adding a regulariser may result in a non-feasible optimisation scheme, due to the computational cost of estimating the (modified) Jacobians. In order to make minimal impact on existing structures, we therefore proceed to linearise the proposed regulariser. This is done by using the alternative formulation in [1], which yields

$$\tilde{\mathcal{R}}_{r_0}(B,C) = \frac{1}{r_0 - \ell}\left(\sum_{i>\ell}\gamma_i(B,C)\right)^2 - \sum_{i>\ell}\gamma_i^2(B,C), \tag{17}$$

where ℓ is the smallest non-negative integer fulfilling $\gamma_\ell \geq \frac{1}{r_0-\ell}\sum_{i>\ell}\gamma_i(B,C) \geq \gamma_{\ell+1}$. After expanding the parenthesis, we arrive at

$$\tilde{\mathcal{R}}_{r_0}(B,C) = \frac{1}{r_0 - \ell}\sum_{i>\ell}\sum_{\substack{j>\ell \\ i\neq j}}\gamma_i\gamma_j - \left(1 - \frac{1}{r_0 - \ell}\right)\sum_{i>\ell}\gamma_i^2 \,. \tag{18}$$

For each $i > \ell$ let $\alpha_i = \sum_{\substack{j>\ell \\ i\neq j}}\gamma_j$, then

$$\tilde{\mathcal{R}}_{r_0}(B,C) = \sum_{i>\ell} f_i(\gamma_i(B,C)), \tag{19}$$

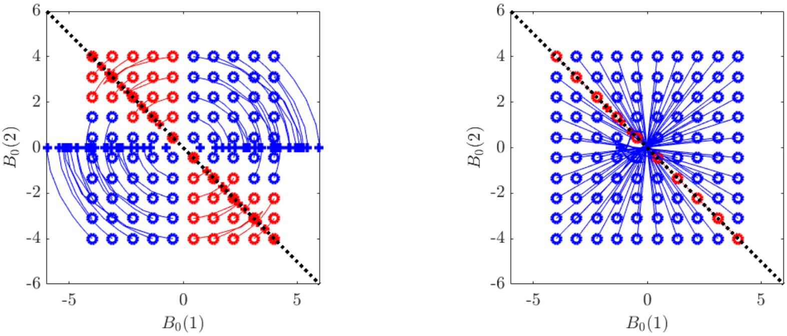


Figure 1: Convergence for different initial values of the first bilinear factor B_0, for regular VarPro (left) and using overparameterisation (right). Circles indicate the initial positions, whereas crosses indicate the final positions. The lines connecting the markers indicate the intermediate updates. The path is coloured blue if the algorithm converged to the correct global minimum for the corresponding initial point, otherwise it is coloured red. The dotted line indicate the fixed points, discussed in Section 6.

where $f_i(x) = \kappa\alpha_i x - (1-\kappa)x^2$ with $\kappa = 1/(r_0 - \ell)$. By considering the first order Taylor expansion $f_i(x) \approx f_i(x_0) + f_i'(x_0)(x - x_0)$ termwise (discarding constant terms) about x_0, we get the approximation

$$\tilde{\mathcal{R}}_{r_0}(B,C) \approx \sum_{i>\ell} w_i^{(t)} \left(\|B_i^{(t)}\|^2 + \|C_i^{(t)}\|^2 \right), \tag{20}$$

where $B^{(t)}$ and $C^{(t)}$ are the current iterates, and

$$w_i^{(t)} = \begin{cases} 0, & i \leq \ell \\ \frac{1}{2} f_i' \left(\frac{\|B_i^{(t)}\|^2 + \|C_i^{(t)}\|^2}{2} \right) & i > \ell\,. \end{cases} \tag{21}$$

When $\operatorname{rank}(BC^T) \leq r_0$, we use the right limit of the weights, *i.e.* $w_i^{(t)} = 0$ for $i \leq r_0$ and $\frac{1}{2} f_{r_0}' \left((\|B_{r_0}^{(t)}\|^2 + \|C_{r_0}^{(t)}\|^2)/2 \right)$ otherwise. We give a more detailed exposition of the algorithm in the supplementary material.

6 Why Overparameterisation?

Consider the original problem of solving (2). One could, with good reason, ask whether it suffices to use VarPro with a bilinear parameterisation admitting a solution of rank at most r_0, *i.e.* with $k = r_0$ columns. Why bother overparameterising with $k > r_0$ columns and regularise the excess columns? In this section we will show, by a toy example, that constraining the solvers to seek solutions in a low rank manifold may converge to incorrect solutions for a significant part of initial points, whereas, by using overparameterisation, convergence to false minima is significantly reduced.

Consider the problem of finding a 2×2 matrix of rank 1, minimising $\|\mathcal{A}(X) - b\|^2$, where

$\mathcal{A}(X) = A \operatorname{vec} X$, with

$$A = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \tag{22}$$

The problem is constructed such that there is a unique solution to $\mathcal{A}(X) = b$, given by

$$X^* = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \tag{23}$$

which is also a rank 1 matrix. Therefore, the global minimum is 0, with the unique minimizer X^*. Note, however, that for regular VarPro, any bilinear factorisation on the form

$$\bar{B} = \frac{\alpha}{\sqrt{2}} \begin{bmatrix} 1 & -1 \end{bmatrix}^T \quad \text{and} \quad \bar{C} = \frac{1}{\sqrt{2}\alpha} \begin{bmatrix} -1 & 1 \end{bmatrix}^T, \tag{24}$$

where $\alpha \neq 0$, gives a fixed point corresponding to $\bar{X} = \bar{B}\bar{C}^T \neq X^*$. Indeed, the Jacobians J_b and J_c are non-zero, however $J_b^T \varepsilon = 0$ and $J_c^T \varepsilon = 0$, where ε is the residual vector. In fact, VarPro converges locally to points on this line, which is shown in Figure 1, where we illustrate the convergence from different initial points. In the example we use $C_0 = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ and fix $\lambda = 1$, but other values result in a similar overall trend. The initial points marked red converge to a point on the line (24), which is a significant part of the initial values. Avoiding such false minima can be alleviated by overparameterisation. Consider using $k = 2$ columns. Starting at the same initial point, by adding a zero column to both B_0 and C_0, it turns out that (24) is not necessarily a fixed point, as $J_b^T \varepsilon$ and $J_c^T \varepsilon$ are non-zero, in general. This is further supported by Figure 1, where it is readily seen that only initial points on the line (24) converges to a non-global minimizer.

7 Non-Rigid Structure From Motion

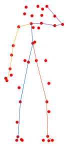
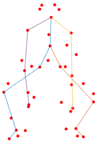
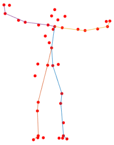
To show the benefits of using the proposed method, we compare it to state-of-the-art methods compatible with (1) on the CMU Motion Capture (MOCAP) dataset. We compare it to the method proposed by [15] which utilises ADMM and the proximal operator of $\mathcal{R}_{r_0}$. We include two relaxations for the "soft-rank" penalty: APGL [26] using nuclear norm, and IRNN [7] which is iteratively reweighted with the robust penalty function MCP [28][1]. Lastly, the standard VarPro [12] method is used in the comparison.

We shall treat this as a low-rank factorisation problem, by employing the approach proposed by Dai *et al.* [9]. Define the matrices

$$X = \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ \vdots \\ X_F \\ Y_F \\ Z_F \end{bmatrix} \quad \text{and} \quad X^\sharp = \begin{bmatrix} X_1 & Y_1 & Z_1 \\ \vdots & \vdots & \vdots \\ X_F & Y_F & Z_F \end{bmatrix}, \tag{25}$$

[1] This was chosen since $f(x) = f_\mu(x) := \mu - \max(\sqrt{\mu} - x, 0)$ (the convex envelope of the "soft rank" objective) is a special case of MCP.

Table 1: Rank vs datafit for the MOCAP experiment. For the bilinear method $k = 2r_0$ columns were used, and was set to a maximum of 80 iterations. Unregularised VarPro and ADMM-$\mathcal{R}_{r_0}$ were allowed to run for the same time (in seconds), or until a local minimum was reached.

	r_0	APGL* [26]	IRNN* [7]	$\mathcal{R}_{r_0}$ [15]	VarPro [12]	Our
Drink 	2	26.4174	**17.0260**	17.0586	**17.0260**	**17.0260**
	3	21.3455	11.0075	12.4201	10.9480	**10.9478**
	4	14.1693	5.9002	6.7927	**5.6855**	**5.6855**
	5	10.3880	4.5602	5.0217	4.4170	**4.3143**
	6	5.9997	4.3858	3.9884	3.5106	**3.4804**
	7	5.7065	3.4676	3.2135	2.9517	**2.9048**
	8	5.4290	2.9927	2.7530	2.6283	**2.4146**
Pickup 	2	46.4320	25.3353	25.4007	**25.3351**	**25.3351**
	3	14.8115	**8.7234**	8.9140	9.3662	**8.7234**
	4	13.7479	6.6112	6.9091	6.5398	**6.4909**
	5	11.1644	**4.8863**	5.4458	4.9002	5.0270
	6	9.5866	3.5880	4.6723	3.5187	**3.4900**
	7	6.3250	3.3919	3.5680	2.7745	**2.7342**
	8	5.8710	2.4373	2.8725	**2.1132**	2.2077
Stretch 	2	31.6038	**21.8045**	21.8224	**21.8045**	**21.8045**
	3	18.4486	**10.6094**	10.6484	11.7657	**10.6094**
	4	14.8941	7.4601	7.4022	7.5132	**7.2913**
	5	11.4202	6.9302	6.1334	5.8904	**5.8798**
	6	9.4485	4.9070	4.9382	**4.6281**	4.6626
	7	8.2575	4.6458	4.1228	3.6614	**3.6249**
	8	6.8711	2.9488	3.1748	2.7256	**2.7044**

(*) APGL and IRNN use the "soft rank" constraint, for which a regularisation parameter must be set. For this experiment, we selected a large range of values and reported the *best values* (not the mean as for the other methods).

where X_i, Y_i and Z_i contain the x-, y- and z-coordinates of the tracked points of the i:th image. Assuming K basis shapes, the matrix $X^\sharp$ may be decomposed into low-rank factors $X^\sharp = CB^\sharp$, where $C \in \mathbb{R}^{F\times K}$ are the shape coefficients and $B^\sharp \in \mathbb{R}^{K\times 3n}$ contains the basis elements. The reason for working with the reshuffled matrices $X^\sharp$ and $B^\sharp$, respectively, is to be able to enforce a stronger rank penalty.

Assuming orthographic cameras, the projection of the scene points are given by $M_i = R_iX_i$, where $R_i \in \mathbb{R}^{2\times 3}$, with $R_iR_i^T = I_2$. A suitable objective function [19] is thus given by

$$\min_{\operatorname{rank}(X^\sharp)\leq K} \|RX - M\|_F^2 + \|DX^\sharp\|_F^2, \tag{26}$$

where $R \in \mathbb{R}^{2F\times 3F}$ is a block-diagonal matrix with the camera matrices R_i on the main diagonal, and $M \in \mathbb{R}^{2F\times n}$ contain the image points. It is known that only considering the datafit as an object is not ideal for NRSfM [9, 19], and, in some cases, promotes non-physical high-rank solutions, and, further penalising the derivative of the 3D projections have been suggested to increase performance [9]. Therefore, the term containing a (modified) difference operator $D : \mathbb{R}^F \to \mathbb{R}^{\lfloor F/2 \rfloor}$, where $\lfloor \cdot \rfloor$ is the floor operator, is included to promote realistic reconstructions. The difference operator is modified to be block-diagonal, hence does not affect the structure of the Jacobians. The results can be seen in Table 1.

Note that we only report the best results for APGL and IRNN as they require the corresponding regularisation parameter to be correctly set. We, therefore, run several tests with varying regularisation strengths and pick the best result for each rank level. This is to minimise potential shrinking bias. In all but three cases the proposed method produced the lowest mean value.

8 Conclusions

In this paper we have presented a novel unification of bilinear parameterisation and rank regularisation utilising overparameterisation to achieve new theoretical optimality guarantees. These results were previously only known in the context of rank penalisation objectives, for which second-order methods are not feasible due to non-differentiability. Using our proposed algorithm we are able to lift essential parts of the theoretical framework developed for regularisation methods, while retaining a differentiable objective suitable for second-order methods. Among the theoretical contributions, we show new strong optimality results under the RIP constraint.

We proposed an algorithm based on VarPro, and show increased performance for difficult objectives for estimating the human pose in a Non-Rigid Structure from Motion framework, compared to state-of-the-art methods.

References

[1] Fredrik Andersson, Marcus Carlsson, and Carl Olsson. Convex envelopes for fixed rank approximation. *Optimization Letters*, pages 1–13, 2017.

[2] Francis R. Bach. Convex relaxations of structured matrix factorizations. *CoRR*, abs/1309.3117, 2013. URL http://arxiv.org/abs/1309.3117.

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

[4] 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.

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

[6] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11:1–11:37, 2011.

[7] Lu Canyi, Jinhui Tang, Shuicheng Yan, and Zhouchen Lin. Nonconvex nonsmooth low-rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25, 10 2015.

[8] Marcus Carlsson, Daniele Gerosa, and Carl Olsson. An unbiased approach to compressed sensing. *arXiv preprint*, arXiv:1806.05283, 2018.

[9] 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.

[10] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference*, 2001.

[11] Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

[12] G. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10 (2):413–432, 1973.

[13] 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.

[14] 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.

[15] Viktor Larsson and Carl Olsson. Convex low rank approximation. *International Journal of Computer Vision*, 120(2):194–214, 2016.

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

[17] 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.

[18] 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, Oct 2017.

[19] 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.

[20] Marcus Valtonen Örnhag, Carl Olsson, and Anders Heyden. Bilinear parameterization for differentiable rank-regularization. *CoRR*, abs/1811.11088, 2018. URL `http://arxiv.org/abs/1811.11088`.

[21] 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.

[22] Samet Oymak, Karthik Mohan, Maryam Fazel, and Babak Hassibi. A simplified approach to recovery conditions for low rank matrices. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 2318–2322, 2011.

[23] Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 65–74, 2017.

[24] 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, August 2010.

[25] 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.

[26] Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6, 09 2010.

[27] 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.

[28] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.