Supplementary Material: 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

A. Proofs

In this section we present the proofs of Theorems 2 and 3. Our analysis will make use of the differentiable objective

$$\mathcal{D}(B,C) := \tilde{\mathcal{R}}(B,C) + \|\mathcal{A}BC^T - b\|^2, \tag{26}$$

the non-convex function

$$\mathcal{N}(X) := \mathcal{R}(X) + \|\mathcal{A}X - b\|^2, \tag{27}$$

and the convex function

$$\mathcal{C}(X) = \mathcal{R}(X) + \|X - Z\|_F^2. \tag{28}$$

We will also use the functions

$$\tilde{G}(B,C) = \tilde{\mathcal{R}}(B,C) + \|BC^T\|_F^2, \tag{29}$$

$$G(X) = \mathcal{R}(X) + \|X\|_F^2, \tag{30}$$

$$H(X) = \|\mathcal{A}X - b\|^2 - \|X\|_F^2. \tag{31}$$

Note that $\mathcal{D}(B,C) = \tilde{G}(B,C) + H(BC^T)$ and $\mathcal{N}(X) = G(X) + H(X)$. Throughout the section we use $f = f_\mu$ with f_μ as in (8) (of the main paper) but for simplicity of notation we will suppress the subscript μ. Furthermore, the subdifferential $\partial G(X)$ of G will be of importance. Let $g(x) = f(|x|) + x^2$. The scalar function g has

$$\partial g(x) = \begin{cases} 2x & |x| \geq \sqrt{\mu} \\ 2\sqrt{\mu}\text{sign}(x) & 0 < |x| \leq \sqrt{\mu} \\ 2\sqrt{\mu}[-1,1] & x = 0 \end{cases}. \tag{32}$$

The following lemma shows how to compute ∂G for the matrix case using ∂g.

Lemma 1. *The subdifferential of $G(X)$ is given by*

$$\partial G(X) = \{U\partial g(\Sigma)V^T + M \ : \ \sigma_1(M) \leq 2\sqrt{\mu}, \\ U^T M = 0 \text{ and } MV^T = 0\} \tag{33}$$

where $X = U\Sigma V^T$ is the SVD and $\partial g(\Sigma)$ is the matrix of same size as Σ with diagonal elements $\partial g(\sigma_i)$.

Next we give the stationary point conditions for $\mathcal{D}$ that are needed for proving Theorem 2.

Lemma 2. *Let $B = U\sqrt{\Sigma}$, $C = V\sqrt{\Sigma}$ and $X = U\Sigma V^T$. If (B,C) is a stationary point of $\mathcal{D}$, then*

$$0 = B\partial G(\Sigma) + \nabla H(BC^T)C, \tag{34}$$

$$0 = \partial G(\Sigma)C^T + B^T\nabla H(BC^T). \tag{35}$$

We are now ready to prove Theorem 2.

Proof of Theorem 2. Let $\bar{X} = \bar{B}\bar{C}^T$, $\tilde{X} = \tilde{B}\tilde{C}^T$ and $\Delta X = \tilde{B}\tilde{C}^T - \bar{B}\bar{C}^T$. We first note that the limit

$$\mathcal{N}'_{\Delta X}(\bar{X}) = \lim_{t \searrow 0} \frac{\mathcal{N}(\bar{X} + t\Delta X) - \mathcal{N}(\bar{X})}{t}, \tag{36}$$

exists since $\mathcal{N}$ is a sum of a finite convex function G and a differentiable function H. Our goal is now to show that the limit is non-negative. Suppose that we can find a factorization $B(t)C(t)^T = \bar{X} + t\Delta X$, such that $\mathcal{R}(\bar{X} + t\Delta X) = \tilde{\mathcal{R}}(B(t),C(t))$, $(B(t),C(t))$ is continuous and $(B(0),C(0)) = (\bar{B},\bar{C})$. Then for small enough t we have

$$\mathcal{N}(\bar{X}+t\Delta X)-\mathcal{N}(\bar{X}) = \mathcal{D}(B(t),C(t))-\mathcal{D}(\bar{B},\bar{C}). \tag{37}$$

This quantity is clearly non-negative since $(\bar{B},\bar{C})$ is a local minimizer of $\mathcal{D}$, which would prove that the limit (36) is non-negative. It is not difficult to see that this can be done when the two matrices $\bar{X}$ and $\tilde{X}$ have singular value decompositions with the same U and V. In what follows we will first show that all other cases can be reduced so that the matrices are of this form. When this is done we proceed to construct the factorization $B(t)C(t)^T$ which completes the proof.

The directional derivatives can be computed using the sub-differential

$$\mathcal{N}'_{\Delta X} = \max_{2Z \in \partial G(\bar{B}\bar{C}^T)} \langle 2Z, \Delta X\rangle + \langle \nabla H(\bar{B}\bar{C}^T), \Delta X\rangle. \tag{38}$$

By Lemma 1, the first term becomes

$$\begin{aligned}\langle U\partial G(\Sigma)V^T + M, \Delta X\rangle &= \langle U\partial G(\Sigma)V^T, \tilde{B}\tilde{C}^T\rangle \\ &+ \langle M, \tilde{B}\tilde{C}^T\rangle \\ &- \langle U\partial G(\Sigma)V^T, \bar{B}\bar{C}^T\rangle.\end{aligned} \tag{39}$$

The columns of $\tilde{B}$ can be written as a linear combination of the columns in $\bar{B}$ and those of a matrix $\bar{B}_\perp$ with at most k columns that are perpendicular to $\bar{B}$. Similarly, the columns of $\tilde{C}$ can be written as a linear combination of the columns in $\bar{C}$ and those of a matrix $\bar{C}_\perp$ with at most k columns that are perpendicular to $\bar{C}$. Therefore, we may write

$$\begin{aligned}\tilde{B}\tilde{C}^T &= \begin{bmatrix}\bar{B} & \bar{B}_\perp\end{bmatrix}\begin{bmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{bmatrix}\begin{bmatrix}\bar{C}^T\\ \bar{C}_\perp^T\end{bmatrix} \\ &= \bar{B}K_{11}C^T + \bar{B}K_{12}\bar{C}_\perp^T \\ &\quad + \bar{B}_\perp K_{21}\bar{C}^T + \bar{B}_\perp K_{22}\bar{C}_\perp^T,\end{aligned} \tag{40}$$

where $\bar{B}^T\bar{B}_\perp = 0$ and $\bar{C}^T\bar{C}_\perp = 0$. Our goal is now to show that the terms K_{12} and K_{21} and the off diagonal elements of K_{11} vanish from (38) and can be assumed to be zero.

For the last term of (39) we have

$$\begin{aligned}\langle U\partial G(\Sigma)V^T, \bar{B}\bar{C}^T\rangle &= \langle \partial G(\Sigma), U^T\bar{B}\bar{C}^T V\rangle \\ &= \langle \partial G(\Sigma), \Sigma\rangle,\end{aligned} \tag{41}$$

which is clearly independent of $\tilde{B}$ and $\tilde{C}$. The first term of (39) reduces to

$$\begin{aligned}\langle U\partial G(\Sigma)V^T, \tilde{B}\tilde{C}^T\rangle &= \langle U\partial G(\Sigma)V^T, \bar{B}K_{11}\bar{C}^T\rangle \\ &= \langle \bar{B}^T U\partial G(\Sigma)V^T\bar{C}, K_{11}\rangle \\ &= \langle \Sigma\partial G(\Sigma), K_{11}\rangle.\end{aligned} \tag{42}$$

Note that the off diagonal elements of K_{11} vanish from this expression since $\Sigma\partial G(\Sigma)$ is diagonal. Similarly, the second term of (39) reduces to

$$\langle M, \tilde{B}\tilde{C}^T\rangle = \langle M, \bar{B}_\perp K_{22}\bar{C}_\perp^T\rangle. \tag{43}$$

We now consider the second term of (38)

$$\begin{aligned}&\langle \nabla H(\bar{B}\bar{C}^T), \Delta X\rangle = \\ &\quad \langle \nabla H(\bar{B}\bar{C}^T), \bar{B}K_{11}\bar{C}^T + \bar{B}K_{12}\bar{C}_\perp^T \\ &\quad + \bar{B}_\perp K_{21}\bar{C}^T + \bar{B}_\perp K_{22}\bar{C}_\perp^T - \bar{B}\bar{C}^T\rangle.\end{aligned} \tag{44}$$

For the first term we have

$$\begin{aligned}\langle \nabla H(\bar{B}\bar{C}^T), \bar{B}K_{11}\bar{C}^T\rangle &= \langle \nabla H(\bar{B}\bar{C}^T)\bar{C}, \bar{B}K_{11}\rangle \\ &= -\langle \bar{B}\partial G(\Sigma), \bar{B}K_{11}\rangle \\ &= -\langle \bar{B}^T\bar{B}\partial G(\Sigma), K_{11}\rangle \\ &= -\langle \Sigma\partial G(\Sigma), K_{11}\rangle.\end{aligned} \tag{45}$$

Again the off diagonal elements of K_{11} vanish. For the second term of (44) we have

$$\begin{aligned}\langle \nabla H(\bar{B}\bar{C}^T), \bar{B}K_{12}C_\perp^T\rangle &= \langle B^T\nabla H(BC^T), K_{12}C_\perp^T\rangle \\ &= -\langle \partial G(\Sigma)\bar{C}^T, K_{12}\bar{C}_\perp\rangle \\ &= -\langle \partial G(\Sigma)\bar{C}^T\bar{C}_\perp, K_{12}\rangle = 0.\end{aligned} \tag{46}$$

Similarly, the third term is $\langle \nabla H(\bar{B}\bar{C}^T), \bar{B}_\perp K_{21}\bar{C}^T\rangle = 0$. Thus

$$\begin{aligned}\langle \nabla H(\bar{B}\bar{C}^T), \Delta X\rangle &= \langle \nabla H(\bar{B}\bar{C}^T), \bar{B}_\perp^T K_{22}\bar{C}_\perp^T\rangle \\ &\quad - \langle \Sigma\partial G(\Sigma), K_{11}\rangle \\ &\quad - \langle \nabla H(\bar{B}\bar{C}^T), \bar{B}\bar{C}^T\rangle.\end{aligned} \tag{47}$$

Summarizing we see that we have now proven that all the terms in (39) are independent of K_{12}, K_{21} as well as the off diagonal terms of K_{11}. They therefore do not affect the value of $\mathcal{N}'_{\Delta X}$ and can be assumed to be zero. We can now write ΔX as

$$\Delta X = \begin{bmatrix}U & U_\perp\end{bmatrix}\begin{bmatrix}(D-I)\Sigma & 0\\ 0 & \tilde{\Sigma}\end{bmatrix}\begin{bmatrix}V^T\\ V_\perp^T\end{bmatrix}, \tag{48}$$

where D are the diagonal elements of K_{11} and $U_\perp\tilde{\Sigma}V_\perp^T$ is the SVD of $\bar{B}_\perp K_{22}\bar{C}_\perp^T$. Note that $U_\perp^T U = 0$ since U and $U_\perp$ span orthogonal subspaces. Similarly $V_\perp^T V = 0$.

We now consider the directional derivative (36) with $\bar{B} = U\sqrt{\Sigma}$, $\bar{C} = V\sqrt{\Sigma}$. It is clear that for small t the matrix $\bar{X} + t\Delta X$ has the singular value decomposition

$$\begin{bmatrix}U & U_\perp\end{bmatrix}\begin{bmatrix}((1-t)I + tD)\Sigma & 0\\ 0 & t\tilde{\Sigma}\end{bmatrix}\begin{bmatrix}V^T\\ V_\perp^T\end{bmatrix}. \tag{49}$$

We now let

$$B(t) = \begin{bmatrix}U & U_\perp\end{bmatrix}\sqrt{\begin{bmatrix}((1-t)I + tD)\Sigma & 0\\ 0 & t\tilde{\Sigma}\end{bmatrix}}, \tag{50}$$

$$C(t) = \begin{bmatrix}V & V_\perp\end{bmatrix}\sqrt{\begin{bmatrix}((1-t)I + tD)\Sigma & 0\\ 0 & t\tilde{\Sigma}\end{bmatrix}}. \tag{51}$$

Then, we clearly have $\tilde{\mathcal{R}}(B(t), C(t)) = \mathcal{R}(X + t\Delta X)$ for small enough t, which completes the proof. □

Next we will prove Theorem 3. Our results build on those of [43] and we remind the reader that we exclusively use $f_\mu(\sigma) = \mu - \max(\sqrt{\mu} - \sigma, 0)^2$ throughout this section, but suppress the subscript μ. We will use the fact that the directional derivatives in a local minimum are non-negative for all low rank directions to show that $(\bar{B}, \bar{C})$ minimizes the non-convex $\mathcal{N}$ over matrices of $\mathrm{rank} < k$ in Theorem 3. For this we will need the following result:

Lemma 3. *If $\bar{X}$ is a solution to $\min_{\mathrm{rank}(X)\le k}\mathcal{C}(X)$ with $\mathrm{rank}(\bar{X}) < k$ and the singular values of Z fulfill $\sigma_i(Z) \notin [(1-\delta_{2k})\sqrt{\mu}, \frac{\sqrt{\mu}}{(1-\delta_{2k})}]$ then $\bar{X}$ also solves $\min_X \mathcal{C}(X)$.*

Proof of Lemma 3. By von Neumann's trace theorem it is easy to see that the problem $\min_{\text{rank}(X)\leq k}\mathcal{C}(X)$ reduces to a minimization over the singular values of X. We should thus find $\sigma_i(X)$ such that

$$\sum_{i=1}^{n}\underbrace{-\max(\sqrt{\mu}-\sigma_i(X),0)^2+(\sigma_i(X)-\sigma_i(Z))^2}_{:=g_i(\sigma_i(X))} \quad (52)$$

is minimized and at most k singular values are non-zero. The unconstrained minimizers of g_i can be written down in closed form: If $0\leq\sqrt{\mu}<\sigma_i(Z)$ then $\sigma_i(X)=\sigma_i(Z)$ is optimal giving $g_i(\sigma_i(X))=0$. If $0\leq\sigma_i(Z)<\sqrt{\mu}$ then $\sigma_i(X)=0$ is optimal giving $g_i(\sigma_i(X))=-\mu+\sigma_i(Z)^2$. Hence for any solution of $\min_{\text{rank}(X)\leq k}\mathcal{C}(X)$ we have $\sigma_i(X)=0$ if $0\leq\sigma_i(Z)\leq\sqrt{\mu}$. There are now two cases:

1. If $\sigma_{k+1}(Z)<\sqrt{\mu}$ then the sequence of unconstrained minimizers has at most k non-zero values. Thus, in this case the resulting X solves both $\min_X\mathcal{C}(X)$ and $\min_{\text{rank}(X)\leq k}\mathcal{C}(X)$.
2. If $\sigma_{k+1}>\sqrt{\mu}$ we will not be able to select $\sigma_i(X)=\sigma_i(Z)$ for all i where $0\leq\sqrt{\mu}<\sigma_i(Z)$. Choosing $\sigma_i(X)=0$ gives $g_i(0)=-\mu+\sigma_i(Z)^2<0$. Since $\sigma_i(Z)$ is decreasing with i it is clear that the smallest value is obtained when selecting $\sigma_i(X)=\sigma_i(Z)$ for $i=1,...,k$.

We now conclude that if $\text{rank}(\bar{X})<k$ then we are in case 1 and therefore $\bar{X}$ solves the unconstrained problem. □

We are now ready to give the proof of Theorem 3.

Proof of Theorem 3. Since $\mathcal{C}$ and $\mathcal{N}$ has the same subdifferential (see [37]) at $\bar{X}=\bar{B}\bar{C}^T$ it is clear that the directional derivatives $\mathcal{C}'_{\Delta X}(\bar{X})=\mathcal{N}'_{\Delta X}(\bar{X})\geq 0$, where $\Delta X=\tilde{X}-\bar{B}\bar{C}^T$ and $\text{rank}(\tilde{X})\leq k$. By convexity of $\mathcal{C}$ it is then also clear that

$$\bar{B}\bar{C}^T\in\operatorname*{arg\,min}_{\text{rank}(X)\leq k}\mathcal{C}(X). \quad (53)$$

Since $\text{rank}(\bar{B}\bar{C}^T)<k$, $\bar{B}\bar{C}^T$ is also the unrestricted global minimizer of $\mathcal{C}(X)$ according to Lemma 3. By Lemma 3.1 of [43] it is then a stationary point of $\mathcal{N}(X)$.

What remains now is to prove that $\bar{X}=\bar{B}\bar{C}^T$ is a global minimizer of $\mathcal{N}$ over all line segments $\bar{X}+t\Delta X$. This can be done by estimating the growth of the directional derivatives along such lines. For this purpose we consider the functions G and H defined as in (30) and (31). Note that $\bar{X}$ is a stationary point of $\mathcal{N}(X)=G(X)+H(X)$ if and only if $-\nabla H(\bar{X})=2Z\in\partial G(\bar{X})$.

Since $\nabla H(\bar{X}+t\Delta X)-\nabla H(\bar{X})=t\nabla H(\Delta X)=2t(\mathcal{A}^*\mathcal{A}\Delta X-\Delta X)$ we have

$$\begin{aligned}\langle\nabla H(\bar{X}+t\Delta X)-\nabla H(\bar{X}),t\Delta X\rangle=\\ 2t^2(\|\mathcal{A}\Delta X\|^2-\|\Delta X\|_F^2),\end{aligned} \quad (54)$$

and due to RIP $\|\mathcal{A}\Delta X\|^2-\|\Delta X\|_F^2\geq-\delta_{2r}\|\Delta X\|^2$. From Corollary 4.2 of [43] we see that for any $2Z'\in\partial G(\bar{X}+t\Delta X)$ we have

$$\langle Z'-Z,t\Delta X\rangle>t^2\delta_{2r}\|\Delta X\|_F^2, \quad (55)$$

as long as $t\neq 0$. Since $G'_{\Delta X}(X)=\max_{2Z\in\partial G(X)}\langle 2Z,\Delta X\rangle$, $H'_{\Delta X}(X)=\langle\nabla H(X),\Delta X\rangle$ and $2Z+\nabla H'(\bar{X})=0$ we get

$$\mathcal{N}'_{\Delta X}(\bar{X}+t\Delta X)\geq\langle 2Z'+\nabla H(\bar{X}+t\Delta X),\Delta X\rangle>0 \quad (56)$$

This shows that $\bar{X}$ solves (9). That $\bar{X}$ also solves (10) is now a consequence of the fact that $\mathcal{R}(X)\leq\mu\text{rank}(X)$ with equality if X have no singular values in the interval $(0,\sqrt{\mu}]$. Note that $\bar{X}$ is the unrestricted minimizer of $\mathcal{C}(X)$, where the singular values of Z fulfill $\sigma_i(Z)\notin\left[(1-\delta_{2k})\sqrt{\mu},\frac{\sqrt{\mu}}{1-\delta_{2k}}\right]$. Since the solution to this problem is hard thresholding $\bar{X}$ has no singular values in $\left(0,\frac{\sqrt{\mu}}{1-\delta_{2k}}\right]\supset(0,\sqrt{\mu}]$. □

For completeness we give the proofs that were previously omitted.

Proof of Lemma 1. With some abuse of notation we define the function $g:\mathbb{R}^n\to\mathbb{R}$ by $g(\mathbf{x})=\sum_{i=1}^n g(x_i)$, where x_i, $i=1,...,n$ are the elements of $\mathbf{x}$ and $g(x)=f(|x|)+x^2$. The function g is an absolutely symmetric convex function and G can be written $G(X)=g\circ\sigma(X)$, where $\sigma(X)$ is the vector of singular values of X. Then according to [39] the matrix $Y\in\partial G(X)$ if and only if $Y=U'\,\text{diag}(\partial g\circ\sigma(X))V'^T$ when $X=U'\,\text{diag}(\sigma(X))V'^T$. (Here we use the full SVD with square orthogonal matrices U' and V'.) Now given a thin SVD $X=U\Sigma V^T$ all possible full SVD's of X can be written

$$X=\begin{bmatrix}U & U_\perp\end{bmatrix}\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}V^T\\ V_\perp^T\end{bmatrix}, \quad (57)$$

where $U_\perp$ and $V_\perp$ are singular vectors corresponding to singular values that are zero. Note that $U_\perp$ and $V_\perp$ are not uniquely defined since their corresponding singular values are all zero. Therefore we get

$$\begin{aligned}Y&=\begin{bmatrix}U' & U_\perp\end{bmatrix}\begin{bmatrix}\partial g(\Sigma) & 0\\ 0 & D\end{bmatrix}\begin{bmatrix}V'^T\\ V_\perp^T\end{bmatrix}\\ &=U'\partial g(\Sigma)V'^T+U_\perp DV_\perp^T,\end{aligned} \quad (58)$$

where D is a diagonal matrix with elements in $2\sqrt{\mu}[-1,1]$. It is clear that $\sigma_1(U_\perp DV_\perp^T)=\sigma_1(D)\leq 2\sqrt{\mu}$. Furthermore, since $U_\perp$ and $V_\perp$ can be any orthogonal bases of the spaces perpendicular to the column and row spaces of X, it is clear that any matrix M fulfilling $U^TM=0$, $MV=0$ and $\sigma_1(M)\leq 2\sqrt{\mu}$ can be written $M=U_\perp DV_\perp^T$, hence

$$\begin{aligned}\partial G(X)=\{U\partial g(\Sigma)V^T+M\ :\ \sigma_1(M)\leq 2\sqrt{\mu},\\ U^TM=0,\ MV=0\}.\end{aligned} \quad (59)$$

□

Proof of Lemma 2. The gradients of $\tilde{G}$ are given by

$$\nabla_B \tilde{G}(B,C) = \nabla_B(\tilde{\mathcal{R}}(B,C)) + \nabla_B(\|BC^T\|_F^2). \quad (60)$$

For the first term we get

$$\nabla_{B_i} \tilde{\mathcal{R}}(B,C) = f'\left(\frac{\|B_i\|^2 + \|C_i\|^2}{2}\right) B_i. \quad (61)$$

With $B = U\sqrt{\Sigma}$ and $C = V\sqrt{\Sigma}$ we get

$$\nabla_B \tilde{\mathcal{R}}(B,C) = B \begin{bmatrix} f'(\sigma_1) & 0 & \dots \\ 0 & f'(\sigma_2) & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} = Bf'(\Sigma), \quad (62)$$

which gives

$$\nabla_B \tilde{G}(B,C) = Bf'(\Sigma) + 2BC^TC = B(f'(\Sigma) + 2\Sigma). \quad (63)$$

For a non-zero σ we have $\partial g(\sigma) = \{f'(\sigma) + 2\sigma\}$ and therefore

$$\nabla_B \tilde{G}(B,C) = B(\partial G(\Sigma)), \quad (64)$$

where $g(X) = \mathcal{R}_\mu(X) + \|X\|_F^2$. Similarly we get

$$\nabla_C \tilde{G}(B,C) = C(\partial G(\Sigma)). \quad (65)$$

If (B,C) is a stationary point then

$$0 = B\partial G(\Sigma) + \nabla H(BC^T)C, \quad (66)$$
$$0 = C\partial G(\Sigma) + (\nabla H(BC^T))^T B. \quad (67)$$

The second equation can be re-written to the form stated in the lemma. □

B. Implementation Details

In this section we present some more details on our Iteratively Reweighted VarPro approach. Recall that our approach consists of three main steps. In the first step we make a quadratic approximation (20) of the regularization term by replacing $\tilde{\mathcal{R}}(B,C)$ with $\sum_{i=1}^k w_i^{(t)} \left(\|B_i\|^2 + \|C_i\|^2\right)$ as described in Section 4.

In the second step we apply one step of VarPro with the Ruhe Wedin approximation, see [33] for details on the implementation. VarPro uses Jacobians with respect to both the B and C parameters. In our case we have two terms that needs to be linearized. The regularization term can be written

$$\|\operatorname{diag}(w^{(t)})B\|_F^2 + \|\operatorname{diag}(w^{(t)})C\|_F^2, \quad (68)$$

where $\operatorname{diag}(w^{(t)})$ is a diagonal matrix with the weights $w_i^{(t)}$ in the diagonal. The residuals $\operatorname{diag}(w^{(t)})B$ are already linear and by column stacking the varaibles we can write them as $J_B^{\text{reg}}\mathbf{b}$, where $\mathbf{b}$ is a column stacked version of B. If B has k columns the matrix J_B^{reg} will consist of k copies of the matrix $\operatorname{diag}(w^{(t)})$. Additionally, each row of J_B^{reg} has only one non-zeros element making the matrix extremely sparse. Similarly, we obtain the contribution due to the second bilinear factor$C$, which can be written as $J_C^{\text{reg}}\mathbf{c}$. Here we use $\mathbf{c} = \operatorname{vec}(C^T)$, as it alleviates the computations of the data terms, hence J_C^{reg} consists of a k copies of $\operatorname{diag}(w^{(t)})$ permuted to match this design choice. Given a current iterate $(\mathbf{b}^{(t)}, \mathbf{c}^{(t)})$ we write the regularization term as $\|J_B^{\text{reg}}\delta\mathbf{b} + \mathbf{r}_B\|^2 + \|J_C^{\text{reg}}\delta\mathbf{c} + \mathbf{r}_C\|^2$, where $\mathbf{r}_B = J_B^{\text{reg}}\mathbf{b}^{(t)}$, $\mathbf{r}_C = J_C^{\text{reg}}\mathbf{c}^{(t)}$, $\mathbf{b} = \mathbf{b}^{(t)} + \delta\mathbf{b}$ and $\mathbf{c} = \mathbf{c}^{(t)} + \delta\mathbf{c}$.

Linearizing the residuals $\mathcal{A}BC^T - b$ around $(\mathbf{b}^{(t)}, \mathbf{c}^{(t)})$ gives an expression of the form

$$J_B^{\text{data}}\delta\mathbf{b} + J_C^{\text{data}}\delta\mathbf{c} + \mathbf{r}^{\text{data}}. \quad (69)$$

The particular shape of the Jacobians in this expression depends on the application; however, in all of our applications they are sparse. For example, in the missing data problem each residual corresponds to an element of the matrix X which in turn only depends on k elements of B and C. Locally we may now write the objective function as

$$\|J_B\delta\mathbf{b} + J_C\delta\mathbf{c} + \mathbf{r}\|^2, \quad (70)$$

where

$$J_B = \begin{bmatrix} J_B^{\text{reg}} \\ 0 \\ J_B^{\text{data}} \end{bmatrix}, \; J_C = \begin{bmatrix} 0 \\ J_C^{\text{reg}} \\ J_C^{\text{data}} \end{bmatrix}, \; \mathbf{r} = \begin{bmatrix} \mathbf{r}_B \\ \mathbf{r}_C \\ \mathbf{r}^{\text{data}} \end{bmatrix}. \quad (71)$$

It was shown in [32] that each step of VarPro is equivalent to first minimizing (70) with the additional dampening term $\lambda\|\delta\mathbf{b}\|^2$ and then performing an exact optimization of (20) over the C-variables (when fixing the B-variables to their new values). Since we also have a reweighing we only do one iteration with VarPro before updating the weights $w^{(t)}$.

The above procedure can return stationary points for which $\tilde{\mathcal{R}}(B,C) > \mathcal{R}(BC^T)$. Our last step is designed to 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 $\tilde{\mathcal{R}}(\bar{B},\bar{C}) = \mathcal{R}(\bar{B}\bar{C}^T)$. Therefore we proceed by refactorizing the current iterate using SVD in each iteration. The detailed steps of the bilinear method are summarized in Algorithm 1.

C. Additional Experiments on Real Data

C.1. pOSE: Psuedo Object Space Error

In this section we compare the energies over time for ADMM optimizing the same energy [37], *i.e.* with the regularizer $\mathcal{R}$, and $f = f_\mu$ as in (8) (of the main paper), and

Input: Robust penalty function f, linear operator $\mathcal{A}$ and regularization parameter μ, damping parameter λ.
Initialize B and C with random entries
while *not converged* **do**
  Compute weights $w^{(t)}$ from current iterate (B, C)
  Compute the vectorizations $\mathbf{b} = \text{vec}(B)$, $\mathbf{c} = \text{vec}(C^T)$
  Compute residuals $\mathbf{r}_B$ $\mathbf{r}_C$, and Jacobians J_B^{data} and J_B^{data} depending on $\mathcal{A}$
  Compute residual $\mathbf{r}^{\text{reg}}$, and Jacobians J_B^{reg} and J_C^{reg}
  Create full residual $\mathbf{r}$ and Jacobians J_B and J_C
  Compute $\tilde{J}^T\tilde{J} + \lambda I = J_B^T(I - J_C J_C^+)J_B + \lambda I$
  Compute $\mathbf{b}' = \mathbf{b} - (\tilde{J}^T\tilde{J} + \lambda I)^{-1} J_B r$ and reshape into matrix B'
  Compute C' by minimizing (20) with fixed B'
  if $\mathcal{R}(B'C'^T) + \|\mathcal{A}(B'C'^T) - b\|^2 < \mathcal{R}(BC^T) + \|\mathcal{A}(BC^T) - b\|^2$ **then**
    $[U, \Sigma, V] = \text{svd}(B'C'^T)$
    Update $B = U\sqrt{\Sigma}$ and $C = V\sqrt{\Sigma}$
    Decrease λ
  else
    Increase λ
  end
end

Algorithm 1: Outline of the bilinear method.

our proposed method. We let the bilinear method run until convergence, and let ADMM execute the same time in seconds. As a comparison we use the nuclear norm relaxation and the discontinuous rank regularization. The results of the experiment are shown in Figure 6.

Again, note that the bilinear method optimizes the same energy as ADMM-$\mathcal{R}_\mu$, and that, despite the initial fast lowering of the objective value, the ADMM approach fails to reach the global optimum, within the allotted 150 seconds. This holds true for all methods employing ADMM. In all experiments, the control parameter $\eta = 0.5$, and the μ parameter was chosen to be smaller than all non-zero singular values of the best known optimum (obtained using VarPro). For a fair comparison, the μ-value for the nuclear norm relaxation, was modified due to the shrinking bias, and was chosen to be the smallest value of μ for which a solution with accurate rank was obtained. Due to this modification, the energy it minimizes is not directly correlated to the others, but is shown for completeness. Furthermore, the iteration speed of ADMM is significantly faster than for VarPro, and therefore we show the elapsed time (in seconds) for all methods. The reported values are averaged over 50 instances with random initialization.

C.2. Background Extraction

The missing data problem formulation can also be used in *e.g.* background extraction, where the goal is to separate

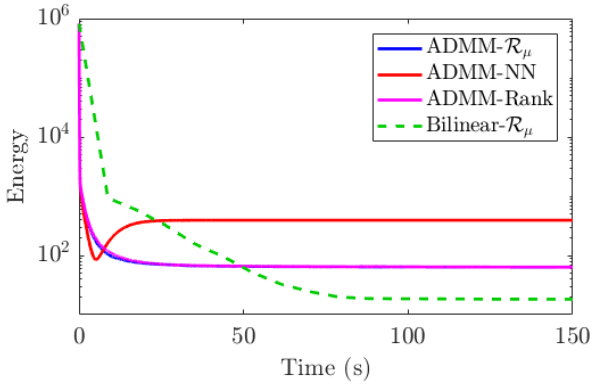

Figure 6. The average energy for the pOSE problem over 50 instances with random initializations, for test sequence *Door*. (Note that the energy for ADMM-Rank and ADMM-$\mathcal{R}_\mu$ are very similar).

the foreground from the background in a video sequence. For this experiment, security footage of an airport is used. The frame size is 144×176 pixels, and we use the first 200 frames, as in [30]. The camera does not move, hence the background is static.

By concatenating the vectorization of the frames into a matrix we expect it to be additively decomposable in terms of a low rank matrix (background) and a sparse matrix (foreground). We follow the setup used in [8], and crop the width to half of the height, and shift it 20 pixels to the right after 100 frames to simulate a virtual pan of the camera. This increases the complexity of the background, as it is no longer static. Lastly, we randomly drop 70 % of the entries. To allow for smaller singular values, we use Geman, as it is a robust penalty with shrinking bias. The results are shown in Figure 8.

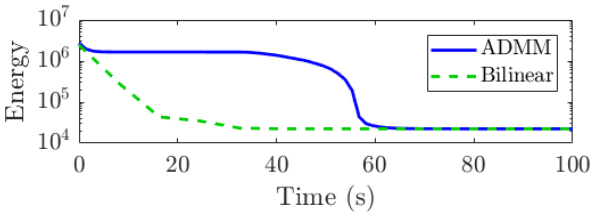

Figure 7. Energy minimization comparison for the background extraction experiment.

Initially ADMM struggles to find the correct balance between lowering the rank and fitting the data, which is seen in Figure 7, where the objective is almost unaffected the first forty seconds. At this point, the bilinear method has already converged.

C.3. Photometric Stereo

Photometric stereo can be used for estimating depth and surface orientation from images of the same object and view

Figure 8. Background extraction using Geman. Samples from frame no. 40, 70, 100, 130, 170 and 200. *Top row:* Original images. *Middle row:* Training data with 70 % missing data. *Bottom row:* Reconstruction of background (bilinear method).

with varying lighting directions. Assuming M lighting directions and N pixels define $I \in \mathbb{R}^{M \times N}$, where I_{ij} is the light intensity for lighting direction i and pixel j. Assuming Lambertian reflectance, uniform albedo and a distant light source, $I = LN$, where $L \in \mathbb{R}^{M \times 3}$ contain the lighting directions and $N \in \mathbb{R}^{3 \times N}$ the unknown surface normals. Thus, the resulting problem is to find a rank 3 approximation of the intensity matrix I.

We use the Harvard Photometric Stereo testset [19], which contains images of various objects from varying lighting direction. The images are scaled to 160×125 pixels, and only the foreground pixels are used in the optimization. Similar to [8], we introduce missing data by thresholding dark pixels with pixel value less than 40 and bright pixels with pixel value more than 205. The measurement matrix is reconstructed using the bilinear method and the ADMM equivalent with the $\mathcal{R}_\mu$ regularization. The result is shown in Figure 9. We let the bilinear method run until convergence and let the ADMM equivalent run for the same time in seconds, at which point the objective value is still decreasing when ADMM is interrupted; however, the reduction is almost negligible. In all cases ADMM fails to converge to a low rank solution in the same time as the bilinear method, which yields a consistent result.

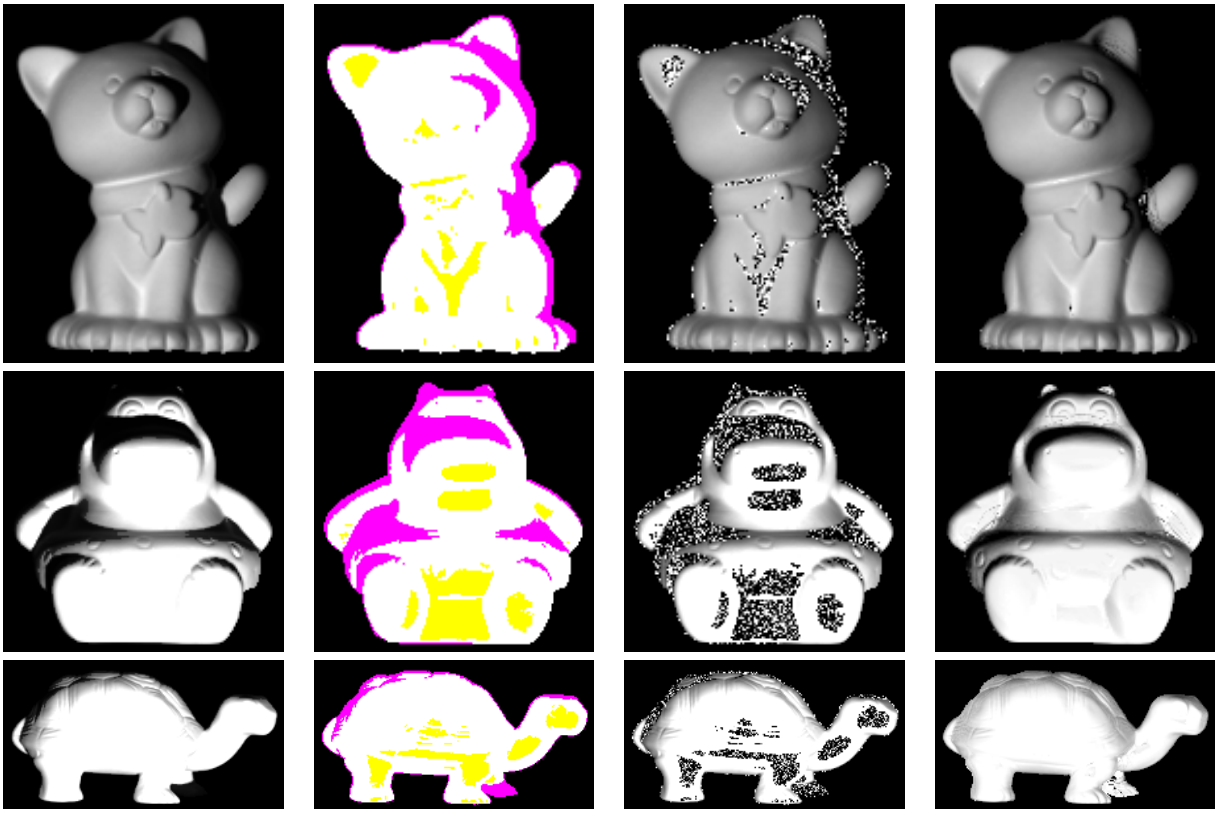

Figure 9. Images from the photometric stereo experiment. *From left to right*: (a) Ground truth image, (b) missing data mask with static background (black), dark pixels (purple), bright pixels (yellow), (c) reconstruction using ADMM, and (d) reconstruction using the Bilinear formulation.

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, 15, 16

[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] R. T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10:439–451, 1988. 16

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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

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

[27] 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

[28] 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

[29] 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

[30] 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. 15

[31] 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

[32] 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, 14

[33] 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, 14

[34] 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

[35] 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

[36] 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

[37] V. Larsson and C. Olsson. Convex low rank approximation. *International Journal of Computer Vision*, 120(2):194–214, 2016. 2, 3, 4, 5, 7, 13, 14

[38] 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

[39] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1):173–183, 1995. 13

[40] 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

[41] 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

[42] 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

[43] 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, 12, 13

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

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

[53] 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

[54] 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

[55] 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

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