Minimizing the Maximal Rank

Bylow, Erik; Olsson, Carl; Kahl, Fredrik; Nilsson, Mikael

Published in:
CVPR 2016, proceedings

2016

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Minimizing the Maximal Rank

Erik Bylow\textsuperscript{1} \hspace{1cm} Carl Olsson\textsuperscript{1} \hspace{1cm} Fredrik Kahl\textsuperscript{1,2} \hspace{1cm} Mikael Nilsson\textsuperscript{1}

\textsuperscript{1} Centre for Mathematical Sciences \hspace{1cm} \textsuperscript{2} Department of Signals and Systems
Lund University, Sweden \hspace{1cm} Chalmers University of Technology, Sweden

\{erikb,calle,micken\}@maths.lth.se \hspace{1cm} fredrik.kahl@chalmers.se

\section*{Abstract}

In computer vision, many problems can be formulated as finding a low rank approximation of a given matrix. Ideally, if all elements of the measurement matrix are available, this is easily solved in the \(L_2\)-norm using factorization. However, in practice this is rarely the case. Lately, this problem has been addressed using different approaches, one is to replace the rank term by the convex nuclear norm, another is to derive the convex envelope of the rank term plus a data term. In the latter case, matrices are divided into sub-matrices and the envelope is computed for each sub-block individually. In this paper a new convex envelope is derived which takes all sub-matrices into account simultaneously. This leads to a simpler formulation, using only one parameter to control the trade-off between rank and data fit, for applications where one seeks low rank approximations of multiple matrices with the same rank. We show in this paper how our general framework can be used for manifold denoising of several images at once, as well as just denoising one image. Experimental comparisons show that our method achieves results similar to state-of-the-art approaches while being applicable for other problems such as linear shape model estimation.

\section{1. Introduction}

Low rank approximation and PCA type procedures are important in many disciplines, for example, statistics, bioinformatics, compression and prediction. In computer vision it has been proven useful for applications such as non-rigid and articulated structure from motion [5, 23, 12], photometric stereo [3], optical flow [13] and linear shape models [8, 22]. The rank of the approximating matrix typically describes the complexity of the solution. Therefore one seeks to find a low rank factorization \(UV^T \approx M\). If the measurement matrix \(M\) is complete and the rank of the approximating matrix is known, then the best approximation, in a least squares sense, can be computed in closed form using the singular value decomposition (SVD) [10].

Alternatively the problem can be formulated as minimization of the objective function

\[ f(X) = \mu \text{rank}(X) + \|X - M\|_F^2. \]  

Here \(\mu\) is a parameter that controls the trade-off between data fit and rank. While the solution is easy to compute using SVD the optimization problem itself is non-convex and non-differentiable. As a consequence it is difficult to modify the formulation without having to resort to heuristic optimization approaches. For example, in case there are missing entries and/or outliers the optimization problem is substantially more difficult. In structure from motion, recent approaches [11, 20] attack these problems by optimizing jointly over fixed size U and V matrices. As a consequence the rank has to be predetermined and the quality of the result is dependent on initialization.

To achieve flexible formulations that are independent of initialization, researchers have instead started to consider convex surrogates of the rank function. Most commonly the convex nuclear norm, or sum-of-singular-values penalty is used [18, 6, 17, 2, 7]. One reason for its popularity is that it can be shown that if the locations of the missing entries are random the approach gives the best low rank approximation [6]. In many computer vision applications missing entry locations are highly correlated which makes the approach break down.

An additional downside of using the nuclear norm is that it has a bias to small solutions. Due to its definition it penalizes both small and large singular values equally. Indeed its proximal operator corresponds to soft thresholding [6]. In contrast, the desirable operation of hard thresholding, which is performed when solving (1) with SVD, leaves the larger singular values unchanged.

A convex formulation that only penalizes the small singular values was recently proposed in [16]. It is shown that the convex envelope of (1) is given by

\[ f^\ast\ast(X) = \sum_{i=1}^n (\mu - \sqrt{\mu - \sigma_i(X)}_+^2) + \|X - M\|_F^2. \]
where $[.]_+$ denotes truncation at 0 and $\sigma_i(X)$, $i = 1, \ldots, n$ are singular values of $X$. Since $f^{**}$ is the convex envelope of $f$ their minimum values coincide and $f^{**}(X)$ is a lower bound $f(X)$ for every $X$. Furthermore, singular values that are larger than $\sqrt{\mu}$ get a constant penalty, which is similar to hard thresholding.

In this paper we are interested in problems where multiple matrices of the same unknown rank need to be estimated. One example where this appears is in manifold estimation. All the tangent spaces of a connected manifold have the same dimension, equal to the dimension of the manifold. Locally a $d$-dimensional manifold can be thought of as a $d$-dimensional tangent space. Therefore approximating the data with a $d$-dimensional manifold can be thought of as locally approximating data with low rank matrices (all of rank $d$). Another problem that can be cast in the same framework is the missing data problem. In [16] it was solved by applying the objective (1) on complete sub-blocks of the measurement matrix. To achieve this, one can divide the data points in a measurement matrix $M$ into different neighborhoods. These neighborhoods form blocks (or sub-matrices) $M_j$ of $M$, see left of Figure 2. Note that in this case, the blocks $M_j$ have the same number of rows as the matrix $M$, but they may vary in sizes and typically have no missing data. We will show how one can compute a low-rank approximation $X$ where all blocks $X_j$ corresponding to $M_j$ have the same rank. These low-rank approximations correspond to the low dimensional (affine) tangent spaces. Further details of the specific formulation for this application will be given in Section 3.1. In this section, we will work with a more general formulation.

We let $\hat{M} = (M_1, M_2, \ldots, M_b)$ be a collection of measurement matrices that we wish to approximate with $\hat{X} = (X_1, X_2, \ldots, X_b)$, where $X_j$, $j = 1, \ldots, b$, are of the same (unknown) rank. Note that the matrices $M_j$ in $\hat{M}$ need not to have the same sizes, and thus $\hat{M}$ should be regarded as a collection of measurement matrices. Our objective function will be of the form

$$\min_{\hat{X}} \mu r(\hat{X}) + \|\hat{X} - \hat{M}\|^2,$$

where the regularization term is

$$r(\hat{X}) = \max(\text{rank}(X_1), \text{rank}(X_2), \ldots, \text{rank}(X_b)),$$

the data fit is measured by

$$\|\hat{X} - \hat{M}\|^2 = \sum_{j=1}^{b} \|X_j - M_j\|^2_F,$$

and $\| \cdot \|_F$ is the regular Frobenious norm. The parameter $\mu$ controls the trade-off between rank and data fit. In practice we are interested in solutions where the ranks of the $X_j$ matrices are the same. It can be seen that the regularizer (4) will achieve this under the assumption that the

2. Regularization With the Maximal Rank

In applications like manifold estimation, one seeks to estimate a manifold by its tangent spaces, where the tangent spaces have a lower dimension than the ambient space, see Figure 1. In particular, all tangent spaces have the same dimension which is also equal to the dimension of the manifold. To achieve this, one can divide the data points in a measurement matrix $M$ into different neighborhoods. These neighborhoods form blocks (or sub-matrices) $M_j$ of $M$, see left of Figure 2. Note that in this case, the blocks $M_j$ have the same number of rows as the matrix $M$, but they may vary in sizes and typically have no missing data. We will show how one can compute a low-rank approximation $X$ where all blocks $X_j$ corresponding to $M_j$ have the same rank. These low-rank approximations correspond to the low dimensional (affine) tangent spaces. Further details of the specific formulation for this application will be given in Section 3.1. In this section, we will work with a more general formulation.

We let $\hat{M} = (M_1, M_2, \ldots, M_b)$ be a collection of measurement matrices that we wish to approximate with $\hat{X} = (X_1, X_2, \ldots, X_b)$, where $X_j$, $j = 1, \ldots, b$, are of the same (unknown) rank. Note that the matrices $M_j$ in $\hat{M}$ need not to have the same sizes, and thus $\hat{M}$ should be regarded as a collection of measurement matrices. Our objective function will be of the form

$$\min_{\hat{X}} \mu r(\hat{X}) + \|\hat{X} - \hat{M}\|^2,$$

where the regularization term is

$$r(\hat{X}) = \max(\text{rank}(X_1), \text{rank}(X_2), \ldots, \text{rank}(X_b)),$$

the data fit is measured by

$$\|\hat{X} - \hat{M}\|^2 = \sum_{j=1}^{b} \|X_j - M_j\|^2_F,$$

and $\| \cdot \|_F$ is the regular Frobenious norm. The parameter $\mu$ controls the trade-off between rank and data fit. In practice we are interested in solutions where the ranks of the $X_j$ matrices are the same. It can be seen that the regularizer (4) will achieve this under the assumption that the
$M_j$ matrices are all of full rank. If for some $j$ we have $r(X) > \text{rank}(X_j)$ then the data term $\|X_j - M_j\|_2^2$ can be reduced by adding another singular value to $X_j$ without affecting any other term.

A common approach would be to simply replace the rank functions in (4) with nuclear norms. However in contrast to the rank function the nuclear norm is not scale invariant. Therefore this will result in a regularizer that penalizes the matrices unevenly. In particular if the matrices $X_j$ have varying sizes. Furthermore, the nuclear norm is only a lower bound on the rank function on the set $\{X; \sigma_1(X) \leq 1\}$, while in contrast our convex envelope will be valid on an unbounded domain. Recall that the convex envelope is by definition the tightest possible lower-bounding convex function, hence the ideal tool for our purposes.

In the following sections we will compute the convex envelope of our formulation via conjugate functions and derive its proximal operator [19].

### 2.1. Conjugate Functions

To find the convex envelope of (3) we consider the conjugate function, which is by definition

$$ f^*(\hat{Y}) = \max_{\hat{X}} \langle \hat{X}, \hat{Y} \rangle - \mu \|\hat{X} - \hat{M}\|_2^2, \quad (6) $$

where $\langle \hat{X}, \hat{Y} \rangle = \sum_{j=1}^b \text{tr}(X_j^T Y_j)$. By completing squares via

$$ \|\hat{X} - (\hat{M} + \frac{\hat{Y}}{2})\|_2^2 = \|\hat{X}\|_2^2 - 2\langle \hat{X}, \hat{M} + \frac{\hat{Y}}{2} \rangle + \|\hat{M} + \frac{\hat{Y}}{2}\|_2^2, \quad (7) $$

the maximization in (6) can be written

$$ \max_k \max_{r(\hat{X}) = k} -\|\hat{X} - \hat{Z}\|_2^2 + \|\hat{Z}\|_2^2 - \|\hat{M}\|_2^2 - \mu k, \quad (8) $$

where $\hat{Z} = \hat{M} + \frac{\hat{Y}}{2}$. For a fixed $k$ the problem is separable in the matrices $X_j$, $j = 1, ..., b$. That is, the optimal $X_j$ can be obtained from the SVD of $Z_j$ giving

$$ X_j = \sum_{i=1}^k \sigma_i(Z_j) u_i v_i^T. \quad (9) $$

Inserting into (8) we get

$$ \max_k -\sum_{i=k+1}^n \|\sigma_i(\hat{Z})\|_2^2 + \|\hat{Z}\|_2^2 - \|\hat{M}\|_2^2 - \mu k. \quad (10) $$

Here $\sigma_i(\hat{Z})$ is the vector $(\sigma_i(Z_1), \sigma_i(Z_2), \ldots, \sigma_i(Z_b))$ and $\| \cdot \|_2$ is the regular euclidean vector norm. To select the maximizing $k$ we note that

$$ \mu k + \sum_{i=k+1}^n \|\sigma_i(\hat{Z})\|_2^2 = \sum_{i=1}^k \mu + \sum_{i=k+1}^n \|\sigma_i(\hat{Z})\|_2^2. \quad (11) $$

Since each entry in the vector $\sigma_i(\hat{Z})$ is positive and decreasing in $i$, its norm $\|\sigma_i(\hat{Z})\|_2^2$ will also be decreasing with $i$. Therefore $k$ should be selected such that

$$ \|\sigma_{k+1}(\hat{Z})\|_2^2 \leq \mu \leq \|\sigma_{k}(\hat{Z})\|_2^2. \quad (12) $$

This gives the conjugate function

$$ f^*(\hat{Y}) = -\sum_{i=1}^n \min(\mu, \|\sigma_i(\hat{Z})\|_2^2) + \|\hat{Z}\|_2^2 - \|\hat{M}\|_2^2. \quad (13) $$

Recall that $\hat{Z}$ depends on $\hat{Y}$ through $\hat{Z} = \hat{M} + \frac{\hat{Y}}{2}$. Next we consider the biconjugate, which is by definition

$$ f^{**}(\tilde{X}) = \max_{\hat{Y}} (\langle \tilde{X}, \hat{Y} \rangle - f^*(\hat{Y})) \quad (14) $$

$$ = \max_{\hat{Y}} \max_{\hat{Z}} \langle \tilde{X}, \hat{Z} \rangle - f^*(\hat{Z}) = \max_\hat{Z} 2\langle \tilde{X}, \hat{Z} - \hat{M} \rangle - f^*(2\hat{Z} - 2\hat{M}). \quad (15) $$

The objective function in (15) can be written

$$ \sum_{i=1}^n \min(\mu, \|\sigma_i(\hat{Z})\|_2^2) - \|\hat{Z} - \hat{M}\|_2^2 + \|\hat{X} - \hat{M}\|_2^2. \quad (16) $$

Using von Neumann’s trace theorem it can be seen that the optimal $Z_j$ has to have an SVD with the same $U$ and $V$ as $X_j$. Therefore the optimization can be reduced to a search over the singular values of the $Z_j$, $j = 1, ..., b$, giving the convex envelope

$$ f^{**}(\tilde{X}) = R_\mu(\tilde{X}) + \|\tilde{X} - \tilde{M}\|_2^2, \quad (17) $$

where

$$ R_\mu(\tilde{X}) = \max_{\hat{Z}} \sum_{i=1}^n \min(\mu, \|\sigma_i(\hat{Z})\|_2^2) - \|\sigma_i(\hat{Z}) - \sigma_i(\tilde{X})\|_2^2. \quad (18) $$

### 2.2. The Proximal Operator of $f^{**}$

The maximization over the singular values in (18) does not seem to have any closed form solution. Evaluation of $f^{**}(\tilde{X})$ therefore has to be done by numerically maximizing the (concave) objective function. At first glance it may therefore seem as though minimization of $f^{**}$ would involve a search over numerical evaluations of $f^{**}$. Fortunately this can be avoided. In this section we show that the proximal operator

$$ \text{prox}_{f^{**}}(\hat{Y}) = \arg \min_{\tilde{X}} f^{**}(\tilde{X}) + \rho \|\tilde{X} - \hat{Y}\|_2^2, \quad (19) $$

which is the basis for ADMM can be computed using a single cone program. The trick is to switch the order of minimization and maximization and thereby obtain a closed form solution for $\tilde{X}$. If $\rho > 0$ the objective function is
closed, proper convex-concave, continuous and the optimization order is therefore justified by the existence of a saddle point, see [19]. To find the optimal $\hat{X}$ we consider the terms of (19) that contain $\hat{X}$

$$-\|\hat{Z} - \hat{X}\|^2 + \|\hat{X} - \hat{M}\|^2 + \rho\|\hat{X} - \hat{Y}\|^2.$$ \hfill (20)

It can be seen (e.g., by taking derivatives of (20)) that the optimal $X$ is given by

$$\hat{X} = \hat{Y} + \frac{\hat{M} - \hat{Z}}{\rho}.$$ \hfill (21)

Inserting into (20) and completing squares gives

$$-\frac{\rho + 1}{\rho}\|\hat{Z} - \hat{W}\|^2 + C,$$ \hfill (22)

where

$$\hat{W} = \frac{\rho\hat{Y} + \hat{M}}{\rho + 1}.$$ \hfill (23)

and

$$C = \frac{2\rho + 1}{\rho}\|\hat{M}\|^2 + \rho\|\hat{Y}\|^2 - \rho\|\hat{Y} + \frac{\hat{M}}{\rho}\|^2.$$ \hfill (24)

Note that $C$ is independent of $\hat{Z}$. In practice we are only interested in finding the optimizers $\hat{Z}$ and $\hat{X}$ and not the objective value itself. Hence we can ignore $C$. We therefore need to maximize

$$\sum_{i=1}^{n} \min(\mu, \|\sigma_i(\hat{Z})\|_2^2) - \frac{\rho + 1}{\rho}\|\hat{Z} - \hat{W}\|^2.$$ \hfill (25)

The terms in the sum only depend on the singular values of the matrices $Z_j$, $j = 1, \ldots, b$. For the second term we have

$$\|\hat{Z} - \hat{W}\|^2 = \|\hat{Z}\|^2 - 2\sum_{j=1}^{b} \langle Z_j, W_j \rangle + \|W\|^2.$$ \hfill (26)

By von Neumann’s trace theorem $\langle Z_j, W_j \rangle \leq \sum_{i=1}^{n} \sigma_i(Z_j) \sigma_i(W_j)$ one sees that the SVD of $Z_j$ has the same $U$ and $V$ as the SVD of $W_j$. Therefore (25) simplifies to

$$\sum_{i=1}^{n} \left( \min(\mu, \|\sigma_i(\hat{Z})\|_2^2) - \frac{\rho + 1}{\rho}\|\sigma_i(\hat{Z}) - \sigma_i(\hat{W})\|_2^2 \right).$$ \hfill (27)

The singular values can now be determined using a cone program. To see this we introduce the auxiliary variables $s_i$, $i = 1, \ldots, n$ and write

$$\max \sum_{i=1}^{n} s_i$$ \hfill (28)

s.t. $s_i \leq \mu - \frac{\rho + 1}{\rho}\|\sigma_i(\hat{Z}) - \sigma_i(\hat{W})\|_2^2$ \hfill (29)

$$s_i \leq \|\sigma_i(\hat{Z})\|_2^2 - \frac{\rho + 1}{\rho}\|\sigma_i(\hat{Z}) - \sigma_i(\hat{W})\|_2^2.$$ \hfill (30)

Note that as we are maximizing the sum of $s_i$, (29) or (30) will always attain equality at the optimal solution. Thus the above program is equivalent to (27). For the singular values to be feasible they have to be decreasing for each block. To enforce this we add linear constraints on the entries of the vectors $\sigma_i(\hat{Z})$ which results in the formulation

$$\max \sum_{i=1}^{n} s_i$$ \hfill (31)

s.t. $\|\sigma_i(\hat{Z}) - \sigma_i(\hat{W})\|_2^2 \leq \frac{\rho}{\rho + 1}(\mu - s_i)$ \hfill (32)

$$\frac{\rho + 1}{\rho}\|\sigma_i(\hat{Z}) - \sigma_i(\hat{W})\|_2^2 \leq - s_i$$ \hfill (33)

$$\sigma_i(\hat{Z}) \geq \sigma_2(\hat{Z}) \geq \ldots \geq \sigma_n(\hat{Z}) \geq 0.$$ \hfill (34)

Equation (32) is easily seen to be convex since the left side is a positive definite quadratic form and the right hand side is linear. To see that the same holds true for (33) we can rewrite this constraint as

$$\|\sigma_i(\hat{Z}) - (\rho + 1)\sigma_i(\hat{W})\|_2^2 \leq \rho \left(\|\sigma_i(\hat{W})\|_2^2 - s_i\right).$$ \hfill (35)

Constraints (32) and (35) can be realized using the cone

$$\{(x_1, x_2, x_3); \ x_1 x_2 \geq \|x_3\|_2^2, \ x_1 + x_2 \geq 0\}.$$ \hfill (36)

This type of cone (which is a rotation of the quadratic cone) is supported in ScDuMi [21] and Mosek [1] which we use to solve (19).

3. Applications

In this section, we present two applications of our framework: (i) Manifold denoising and (ii) Linear shape basis models.

3.1. Manifold Denoising

Manifold denoising can be formulated as seeking affine tangent spaces with the same dimension. If we have a set of images, possibly corrupted with noise, then the assumption is that the true uncorrupted images lie on a low-dimensional manifold. The images, represented by $m_i$, $i = 1, \ldots, N$, are assumed to be column-stacked so one image lies in $\mathbb{R}^n$, where $n$ is the number of pixels. The assumption means that several points which are close to each other should be close to the tangent space of the manifold as illustrated in Figure 1. To determine neighbourhoods, we find for each image point its $K$-closest neighbors in the euclidean distance and consider them to be one block.

Given a set of images, stacked in a measurement matrix $M = [m_1, m_2, \ldots, m_N]$, we determine a collection of blocks via the neighbourhoods, $M = (M_1, M_2, \ldots, M_b)$, see left of Figure 2. Since the images are corrupted by noise,
Table 1: The PSNR using different methods for denoising on the USPS Digits.

<table>
<thead>
<tr>
<th>Method</th>
<th>Input PSNR</th>
<th>Output PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>10.4553</td>
<td><strong>17.5231</strong></td>
</tr>
<tr>
<td>Manifold Denoising</td>
<td>10.4553</td>
<td>15.6656</td>
</tr>
</tbody>
</table>

Each sub-matrix $M_i$ will have high rank, and the task is to find a low-rank approximation $X_i$ for each $M_i$. Note that the different $X_i$ share common variables. Also, since we are interested in the affine tangent spaces, which do not necessarily go through the origin, we add the row-wise mean vector $\bar{x}_i$ of each $X_i$. Assuming that $X_i$ have zero row means, that is $X_i\mathbb{1} = 0$, the fitting terms in the objective function can be written

$$\sum_{i=1}^b \|X_i + \bar{x}_i\mathbb{1}^T - M_i\|^2 = \sum_{i=1}^b \|X_i - (M_i - \bar{m}_i\mathbb{1}^T)\|^2_F + k_i\|\bar{x}_i - \bar{m}_i\|^2_2,$$

where $\bar{m}_i$ is the row mean of $M_i$ and $k_i$ is equal to the number of columns in block $M_i$. To ensure consistency between shared variables, we penalise the differences by adding to the objective

$$\alpha \sum_{i=1}^b \|\mathcal{P}_i(X) - (X_i + \bar{x}_i\mathbb{1}^T)\|^2_F,$$

where $\alpha$ is a weighting factor, $X$ is the approximation of the measurement matrix $M$ and $\mathcal{P}_i(X)$ retrieves block $i$ in $X$.

In summary, we have the following optimization problem

$$\min_{X, \bar{x}, z} \rho\|\hat{X} - \bar{Z} + \hat{\Lambda}\|^2_F +$$

$$\sum_{i=1}^b \left(k_i\|\bar{x}_i - \bar{m}_i\|^2_2 + \alpha\|\mathcal{P}_i(X) - (X_i + \bar{x}_i\mathbb{1}^T)\|^2_F\right)$$

s.t. $X_i\mathbb{1} = 0, \quad i = 1, \ldots, b.$

We have already derived the convex envelope for the terms on the first row (39) and the terms on the second row (40) are convex from the start. To minimize (39) we use the convex envelope and introduce auxiliary variables $Z_i$, which results in

$$\min_{\hat{X}, X, \bar{x}, z} f^\ast\ast(\hat{X}) +$$

$$\sum_{i=1}^b \left(k_i\|\bar{x}_i - \bar{m}_i\|^2_2 + \alpha\|\mathcal{P}_i(X) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F\right)$$

s.t. $X_i = Z_i, \quad Z_i\mathbb{1} = 0, \quad i = 1, \ldots, b,$

and in turn, this leads to the ADMM formulation

$$\min_{\hat{X}, X, \bar{x}, z} f^\ast\ast(\hat{X}) + \rho\|\hat{X} - \bar{Z} + \hat{\Lambda}\|^2_F +$$

$$\sum_{i=1}^b \left(k_i\|\bar{x}_i - \bar{m}_i\|^2_2 + \alpha\|\mathcal{P}_i(X) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F\right)$$

s.t. $Z_i\mathbb{1} = 0, \quad i = 1, \ldots, b.$

We get one part which depends on $\hat{X}$,

$$\min_X f^\ast\ast(\hat{X}) + \rho\|\hat{X} - \bar{Z} + \hat{\Lambda}\|^2_F,$$

which is precisely the proximal operator we have seen before. To minimize with respect to the other variables $\bar{Z}$, $\bar{x}_i$ and $X$ is now straightforward. Keeping the other variables fixed and solving for one we get the following updates:

$$X^{t+1} = \arg\min_X \alpha \sum_{i=1}^b \|\mathcal{P}_i(X) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F +$$

$$\sum_{i=1}^b \left(k_i\|\bar{x}_i - \bar{m}_i\|^2_2 + \alpha\|\mathcal{P}_i(X) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F\right),$$

which is a separable least squares problem,

$$Z_i^{t+1} = \arg\min_{Z_i} \alpha\|\mathcal{P}_i(X^{t+1}) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F +$$

$$\rho\|X_i^{t+1} - Z_i + \hat{\Lambda}_i\|^2_F, \quad i = 1, \ldots, b,$$

since all blocks are independent in the minimization,

$$\bar{x}_i^{t+1} = \arg\min_{\bar{x}_i} \sum_{i=1}^b \left(k_i\|\bar{x}_i - \bar{m}_i\|^2_2 + \right.$$

$$\alpha\|\mathcal{P}_i(X^{t+1}) - Z_i - \bar{x}_i\mathbb{1}^T\|^2_F),$$

$$\hat{X}^{t+1} = \text{prox}_{\rho f^\ast\ast}(\hat{Z}^{t+1} - \hat{\Lambda}^{t+1}).$$

For $\hat{\Lambda}$ we take a step in the ascent direction, that is,

$$\hat{\Lambda}_i^{t+1} = \hat{\Lambda}_i^{t} + X_i^{t+1} - Z_i^{t+1},$$

since again, each block is separable.

Experimental results: USPS. To test the denoising method, we use the USPS dataset [15] of handwritten digits. We choose 100 images of each digit and rescale the intensities to lie between $[0, 1]$. The images are perturbed by Gaussian noise with standard deviation $\sigma = 0.3$. We then stack all images into one measurement matrix $M$, and find the $K = 30$ closest neighbours for each image. With this data, we apply our optimization to obtain the new approximate
matrix $X$ of $M$. Each column in $X$ contains a denoised image corresponding to the noisy image in the same column in $M$.

For comparison we use the well-known work called Manifold Denoising by [14]. This work uses a different approach where a partial differential equation is solved on a graph created by the data points to obtain a manifold.

The results shown in Table 1 where obtained when adding noise with standard deviation $\sigma = 0.3$ to the USPS dataset. For Manifold Denoising we set the number of neighbors to 6 and re-weighting parameter $\lambda = 1$ and a symmetric graph, since that gave the best results in our experiment.

**Experimental results: Single image denoising.** Our method for manifold denoising can also be used to denoise a single image. To apply our method, the image is first divided into several patches, and each patch is considered to be one point in $\mathbb{R}^p$. As above all points, or patches, are then stacked into one measurement matrix $M$. Thereafter,
Table 2: Denoising results from the Cameraman and Lena. BM3D gives a higher PSNR for Lena, but we do better on the Cameraman.

Table 3: The error $\sum_{i=1}^{b} ||X_i - M_i||_F^2$ for the method in [16] and our method. Note that the method in [16] outperforms the nuclear norm relaxation for the same error metric.

3.2. Linear Shape Basis Models

Another application we test our framework on is estimation of linear shape models. A common assumption is that a set of tracked image points moving non-rigidly can be described with a small number of basis elements in each frame. If we let $M_f = (m_{f1}, m_{f2}, \ldots, m_{fN})$ denote the $N$ tracked 2D-points in frame $f$, we want to find shape basis models $(S_1, S_2, \ldots, S_K)$ — each of size $2 \times N$ — and scalar coefficients $C_{f1}, C_{f2}, \ldots, C_{fK}$ such that the points $M_f$ can be described by

$$M_f = \sum_{k=1}^{K} C_{fk} S_k. \quad (52)$$

Stacking the $N$ points in $F$ frames yields a $2F \times N$ measurement matrix $M$. Since we want to use as few basis elements as possible, the matrix $M$ should be of low rank. Due to occlusion and tracking failures, not all points will be seen in all frames. This gives a measurement matrix $M$ with missing data. To handle this we create sub-blocks $M_i$ of $M$, where each $M_i$ has no missing entries, see right of Figure 2. Hence, we have turned the problem into finding low-rank approximations $X_i$ of $M$, where the blocks in $X_i$ share common variables. The objective function we seek to minimize is

$$\min_{\hat{X}, X} \quad f^{**}(\hat{X}) \quad (53)$$

$$\text{s.t.} \quad P_i(X) = X_i \quad i = 1, \ldots, b,$$

where $\hat{X}$ is the collection of blocks $(X_1, X_2, \ldots, X_b)$ and $P_i(X)$ retrieves block $i$ from $X$. The constraint comes from the requirement that the blocks we optimize over shall coincide on the overlap. To minimize this, we use ADMM and our augmented Lagrangian becomes

$$f^{**}(\hat{X}) + \rho ||\hat{X} - \hat{P}(X) + \hat{\Lambda}||^2 - \rho ||\hat{\Lambda}||^2, \quad (54)$$

where $\hat{P}(X) = (P_1(X), \ldots, P_b(X))$.

In each iteration we perform the following updates:

$$\hat{X}_{t+1} = \arg \min_{\hat{X}, t} \quad f^{**}(\hat{X}^t) + \rho ||\hat{X}^t - \hat{P}(X^t) + \hat{\Lambda}^t||^2 \quad (55)$$

$$X_{t+1} = \arg \min_{X} \quad \rho ||\hat{X}^t - \hat{P}(X^t) + \hat{\Lambda}^t||^2 \quad (56)$$

$$\Lambda_{t+1}^i = \Lambda_{t}^i + X_{t+1} - \hat{P}_i(X_{t+1}). \quad (57)$$
When the low-rank approximation is found where only known data has been used, we complete the missing parts in \( X \) by applying the same method as in [16].

**Experimental results.** Our framework has been applied to a number of image sequences obtained from the authors of [16]. The results of the Hand-, Book- and Banner sequences are shown in Figures 7, 8 and 9.

One sees clearly in all sequences that the red and blue points, which are the reconstructed points, obey a motion which is reasonable compared to the input data, green points. The blue points are the reconstructed points which we could track and the red points are the reconstructed positions of points with no measurements available. The found rank for the solution in the Hand sequence is 5, in the Book sequence we get rank 3 and in the Banner sequence we get rank 9. The number of blocks in Hand, Book and Banner was 5, 3 and 19.

To compare with [16], we test their method on the same datasets and measure the error \( \sum_{i=1}^{b} \| M_i - X_i \|_F^2 \) and the results are shown in Table 3. We choose to measure the error on the blocks since that will show if our method differs from [16]. Note that this method was shown to perform better than the nuclear norm relaxation for this application.

As the results in Table 3 show, we do at least as good as they do on these datasets. We investigated the approximated sub-matrices from [16] individually and saw that some sub-matrices had rank 8 and some rank 10. This shows that it is easier to get uniform rank on all sub-matrices using our formulation.

**4. Conclusion**

In this paper we have derived a novel and general convex framework to approximate low-rank matrices. Our method is suitable in situations where several matrices of the same rank need to be approximated. Our main contribution is the derivation of a strong convex formulation that can be optimized in general frameworks using the proximal operator in an ADMM fashion [4]. One of the advantages of our formulation is that there is a single tuning parameter controlling the trade-off between rank and model fit which is important in manifold estimation where the number of sub-matrices may be well above 10,000.

Experimental evaluations showed that our method achieves results similar to state-of-the-art approaches on manifold denoising problems and linear shape basis estimation. It should be mentioned that in the case of manifold denoising neighborhoods are computed using noisy patches. Therefore results could potentially be improved by reestimating neighborhoods using cleaned versions. However, to focus the evaluation on our convex formation we have refrained from such heuristics.

**5. Acknowledgements**

This work has been funded by the Swedish Research Council (grant no. 2012-4213), the Crafoord Foundation and Scientific Research Council, project no. 2012-4215.

**References**


