NON-LINEAR DIMENSIONALITY REGULARIZER FOR SOLVING INVERSE PROBLEMS

Ravi Garg
University of Adelaide
`ravi.garg@adelaide.edu.au`

Anders Eriksson
Queensland University of Technology
`anders.eriksson@qut.edu.au`

Ian Reid
University of Adelaide
`ian.reid@adelaide.edu.au`

ABSTRACT

Consider an ill-posed inverse problem of estimating causal factors from observations, one of which is known to lie near some (unknown) low-dimensional, non-linear manifold expressed by a predefined Mercer-kernel. Solving this problem requires simultaneous estimation of these factors and learning the low-dimensional representation for them. In this work, we introduce a novel non-linear dimensionality regularization technique for solving such problems without pre-training.

We re-formulate Kernel-PCA as an energy minimization problem in which low dimensionality constraints are introduced as regularization terms in the energy. To the best of our knowledge, ours is the first attempt to create a dimensionality regularizer in the KPCA framework. Our approach relies on robustly penalizing the rank of the recovered factors directly in the implicit feature space to create their low-dimensional approximations in closed form.

Our approach performs robust KPCA in the presence of missing data and noise. We demonstrate state-of-the-art results on predicting missing entries in the standard oil flow dataset. Additionally, we evaluate our method on the challenging problem of Non-Rigid Structure from Motion and our approach delivers promising results on CMU mocap dataset despite the presence of significant occlusions and noise.

1 INTRODUCTION

Dimensionality reduction techniques are widely used in data modeling, visualization and unsupervised learning. Principal component analysis (PCAJolliffe (2002)), Kernel PCA (KPCASchölkopf et al. (1998)) and Latent Variable Models (LVMsLawrence (2005)) are some of the well known techniques used to create low dimensional representations of the given data while preserving its significant information.

One key deployment of low-dimensional modeling occurs in solving ill-posed inference problems. Assuming the valid solutions to the problem lie near a low-dimensional manifold (i.e. can be parametrized with a reduced set of variables) allows for a tractable inference for otherwise under-constrained problems. After the seminal work of Candès & Recht (2009); Recht et al. (2010) on guaranteed rank minimization of the matrix via trace norm heuristics Fazel (2002), many ill-posed computer vision problems have been tackled by using the trace norm — a convex surrogate of the rank function — as a regularization term in an energy minimization frameworkCandès & Recht (2009); Zhou et al. (2014). The flexible and easy integration of low-rank priors is one of key factors for versatility and success of many algorithms. For example, pre-trained active appearance models Cootes et al. (2001) or 3D morphable models Blanz & Vetter (1999) are converted to robust feature tracking Poling et al. (2014), dense registration Garg et al. (2013b) and vivid reconstructions of natural videos Garg et al. (2013a) with no *a priori* knowledge of the scene. Various bilinear factorization problems like background modeling, structure from motion or photometric stereo are also addressed with a variational formulation of the trace norm regularization Cabral et al. (2013).

On the other hand, although many non-linear dimensionality reduction techniques — in particular KPCA — have been shown to outperform their linear counterparts for many data modeling tasks, they are seldom used to solve inverse problems without using a training phase. A general (discriminative) framework for using non-linear dimensionality reduction is: (i) learn a low-dimensional representation for the data using training examples via the kernel trick (ii) project the test examples on the learned manifold and finally (iii) find a data point (pre-image) corresponding to each projection in the input space.

This setup has two major disadvantages. Firstly, many problems of interest come with corrupted observations — noise, missing data and outliers — which violate the low-dimensional modeling assumption.Secondly, computing the pre-image of any point in the low dimensional feature subspace is non-trivial: the pre-image for many points in the low dimensional space might not even exist because the non linear feature mapping function used for mapping the data from input space to the feature space is non-surjective.

Previously, extensions to KPCA like Robust KPCA (RKPCANguyen & De la Torre (2009)) and probabilistic KPCA (PKPCASanguinetti & Lawrence (2006)) with missing data have been proposed to address the first concern, while various additional regularizers have been used to estimate the pre-image robustly Bakir et al. (2004); Mika et al. (1998); Kwok & Tsang (2004); Abrahamsen & Hansen (2009).

Generative models like LVMs Lawrence (2005) are often used for inference by searching the low-dimensional latent space for a location which maximizes the likelihood of the observations. Problems like segmentation, tracking and semantic 3D reconstruction Prisacariu & Reid (2011); Dame et al. (2013) greatly benefit from using LVM. However, the latent space is learned *a priori* with clean training data in all these approaches.

Almost all non-linear dimensionality reduction techniques are non-trivial to generalize for solving ill-posed problems (See section 4.2) without a pre-training stage. Badly under-constrained problems require the low-dimensional constraints even for finding an initial solution, eliminating applicability of the standard "projection + pre-image estimation" paradigm. This hinders the utility of non-linear dimensionality reduction and a suitable regularization technique to penalize the non-linear dimensionality is desirable.

Sum and Substance: A closer look at most non-linear dimensionality reduction techniques reveals that they rely upon a non-linear mapping function which maps the data from input space to a (usually) higher dimensional feature space. In this feature space the data is assumed to lie on a low-dimensional hyperplane — thus, *linear low-rank prior is apt in the feature space*. Armed with this simple observation, our aim is to focus on incorporating the advances made in linear dimensionality reduction techniques to their non-linear counterparts, while addressing the problems described above. Figure 1 explains this central idea and proposed dimensionality regularizer in a nutshell with Non Rigid Structure from Motion (NRSfM) as the example application.

Our Contribution: In this work we propose a unified for simultaneous robust KPCA and pre-image estimation while solving an ill-posed inference problem without a pre-training stage.

In particular we propose a novel robust energy minimization algorithm which handles the implicitness of the feature space to directly penalize its rank by iteratively: (i) creating robust low-dimensional representation for the
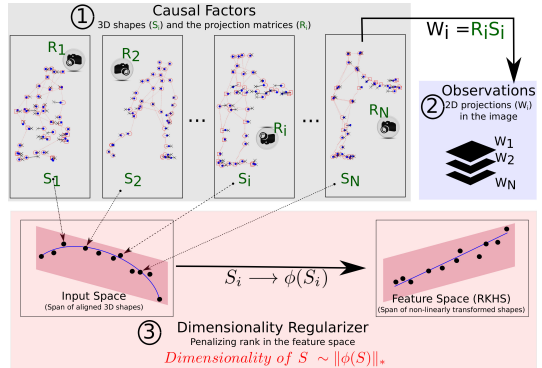
Figure 1: Non-linear dimensionality regularizer for NRSfM. The top part of the figure explains the ill-posed *inverse problem* of recovering the *causal factors (1)*; projection matrices R_i and 3D structures S_i, from 2D image *observations (2)* W_i's, by minimizing the *image reprojection* error $f(W, R, S) = \sum_i \|W_i - R_i S_i\|^2$. Assuming that the recovered 3D structures (S_i's) lies near an unknown non-linear manifold (represented by the blue curve) in the input space, we propose to *regularize the dimensionality of this manifold (3)* — span of the non-linearly transformed shape vectors $\phi(S_i)$'s — by minimizing $\|\phi(S)\|_*$. The non-linear transformation ϕ is defined implicitly with a Mercer kernel and maps the non-linear manifold to a linear low rank subspace (shown in blue line) of RKHS.

data given the kernel matrix in closed form and (ii) reconstructing the noise-free version of the data (pre-image of the features space projections) using the estimated low-dimensional representations in a unified framework.

The proposed algorithm: (i) provides a novel closed form solution to robust KPCA; (ii) yields state-of-the-art results on missing data prediction for the well-known oil flow dataset; (iii) outperforms state-of-the-art linear dimensionality (rank) regularizers to solve NRSfM; and (iv) can be trivially generalized to incorporate other cost functions in an energy minimization framework to solve various ill-posed inference problems.

2 PROBLEM FORMULATION

This paper focuses on solving a generic inverse problem of recovering causal factor $S = [s_1, \ s_2, \ \cdots s_N] \in \mathcal{X} \times N$ from N observations $W = [w_1, \ w_2, \ \cdots w_N] \in \mathcal{Y} \times N$ such that $f(W, S) = 0$. Here function f(observation,variable), is a generic loss function which aligns the observations W with the variable S (possibly via other causal factors. e.g. R or Z in Section 4.1 and 4.2).

If, $f(W, S) = 0$ is ill-conditioned (for example when $\mathcal{Y} \ll \mathcal{X}$), we want to recover matrix S under the assumption that the columns of it lie near a low-dimensional non-linear manifold. This can be done by solving a constrained optimization problem of the following form:

$$\min_{S} \quad rank(\Phi(S))$$
$$s.t. \ f(W, S) \leq \epsilon \tag{1}$$

where $\Phi(S) = [\phi(s_1), \ \phi(s_2), \ \cdots, \ \phi(s_N)] \in \mathcal{H} \times N$ is the non-linear mapping of matrix S from the input space $\mathcal{X}$ to the feature space $\mathcal{H}$ (also commonly referred as Reproducing Kernel Hilbert Space), via a non-linear mapping function $\phi : \mathcal{X} \to \mathcal{H}$ associated with a Mercer kernel K such that $K(S)_{i,j} = \phi(s_i)^T \phi(s_j)$.

In this paper we present a novel energy minimization framework to solve problems of the general form (1).

As our first contribution, we relax the problem (1) by using the trace norm of $\Phi(S)$ — the convex surrogate of rank function — as a penalization function. The trace norm $\|M\|_* =: \sum_i \lambda_i(M)$ of a matrix M is the sum of its eigenvalues $\lambda_i(M)$ and was proposed as a tight convex relaxation[1] of the $rank(M)$ and is used in many vision problems as a rank regularizer Fazel (2002). Although the rank minimization via trace norm relaxation does not lead to a convex problem in presence of a non-linear kernel function, we show in 3.2 that it leads to a closed-form solution to denoising a kernel matrix via penalizing the rank of recovered data (S) directly in the feature space.

With these changes we can rewrite (1) as:

$$\min_{S} \quad f(W, S) + \tau \|\Phi(S)\|_* \tag{2}$$

where τ is a regularization strength.[2]

It is important to notice that although the rank of the kernel matrix $K(S)$ is equal to the rank of $\Phi(S)$, $\|K(S)\|_*$ is merely $\|\Phi(S)\|_{\mathcal{F}}^2$. Thus, directly penalizing the sum of the singular values of $K(S)$ will not encourage low-rank in the feature space.[3]

Although we have relaxed the non-convex rank function, (2) is in general difficult to minimize due to the implicitness of the feature space. Most widely used kernel functions like RBF do not have a explicit definition of the function ϕ. Moreover, the feature space for many kernels is high- (possibly infinite-) dimensional, leading to intractability. These issues are identified as the main

---

[1]More precisely, $\|M\|_*$ was shown to be the tight convex envelope of $rank(M)/\|M\|_s$, where $\|M\|_s$ represent spectral norm of M.

[2]$1/\tau$ can also be viewed as Lagrange multiplier to the constraints in (1).

[3]Although it is clear that relaxing the rank of kernel matrix to $\|K(S)\|_*$ is suboptimal, works like Huang et al. (2012); Cabral et al. (2013) with a variational definition of nuclear norm, allude to the possibility of kernelization. Further investigation is required to compare this counterpart to our tighter relaxation.

barriers to robust KPCA and pre-image estimation Nguyen & De la Torre (2009). Thus, we have to reformulate (2) by applying kernel trick where the cost function (2) can be expressed in terms of the kernel function alone.

The key insight here is that under the assumption that kernel matrix $K(S)$ is positive semidefinite, we can factorize it as: $K(S) = C^T C$. Although, this factorization is non-unique, it is trivial to show the following:

$$\sqrt{\lambda_i(K(S))} = \lambda_i(C) = \lambda_i(\Phi(S))$$

$$\text{Thus:} \quad \|C\|_* = \|\Phi(S)\|_* \quad \forall \, C : C^T C = K(S) \tag{3}$$

where $\lambda_i(.)$ is the function mapping the input matrix to its ith largest eigenvalue.

The row space of matrix C in (3) can be seen to span the eigenvectors associated with the kernel matrix $K(S)$ — hence the principal components of the non-linear manifold we want to estimate.

Using (3), problem (2) can finally be written as:

$$\min_{S,C} \; f(W, S) + \tau \|C\|_*$$

$$s.t. \quad K(S) = C^T C \tag{4}$$

The above minimization can be solved with a soft relaxation of the manifold constraint by assuming that the columns of S lie near the non-linear manifold.

$$\min_{S,C} \; f(W, S) + \frac{\rho}{2} \|K(S) - C^T C\|_{\mathcal{F}}^2 + \tau \|C\|_* \tag{5}$$

As $\rho \to \infty$, the optimum of (5) approaches the optimum of (4) . A local optimum of (4) can be achieved using the penalty method of Nocedal & Wright (2006) by optimizing (5) while iteratively increasing ρ as explained in Section 3.

Before moving on, we would like to discuss some alternative interpretations of (5) and its relationship to previous work – in particular LVMs. Intuitively, we can also interpret (5) from the probabilistic viewpoint as commonly used in latent variable model based approaches to define kernel function Lawrence (2005). For example a RBF kernel with additive Gaussian noise and inverse width γ can be defined as: $K(S)_{i,j} = e^{-\gamma \|s_i - s_j\|^2} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma)$. In other words, with a finite ρ, our *model allows the data points to lie near a non-linear low-rank manifold* instead of on it. Its worth noting here that like LVMs, our energy formulation also attempts to maximize the likelihood of regenerating the training data W, (by choosing $f(W, S)$ to be a simple least squares cost) while doing dimensionality reduction.

Note that in closely related work Geiger et al. (2009), continuous rank penalization (with a logarithmic prior) has also been used for robust probabilistic non-linear dimensionality reduction and model selection in LVM framework. However, unlike Geiger et al. (2009); Lawrence (2005) where the non-linearities are modeled in latent space (of predefined dimensionality), our approach directly penalizes the non-linear dimensionality of data in a KPCA framework and is applicable to solve inverse problems without pre-training.

3 OPTIMIZATION

We approach the optimization of (5) by solving the following two sub-problems in alternation:

$$\min_{S} \quad f(W, S) + \frac{\rho}{2} \|K(S) - C^T C\|_{\mathcal{F}}^2 \tag{6}$$

$$\min_{C} \quad \tau \|C\|_* + \frac{\rho}{2} \|K(S) - C^T C\|_{\mathcal{F}}^2 \tag{7}$$

Algorithm 1 outlines the approach and we give a detailed description and interpretations of both sub-problems (7) and (6) in next two sections of the paper.

3.1 PRE-IMAGE ESTIMATION TO SOLVE INVERSE PROBLEM.

Subproblem (6) can be seen as a generalized pre-image estimation problem: we seek the factor s_i, which is the pre-image of the projection of $\phi(s_i)$ onto the principle subspace of the RKHS stored in

---

Algorithm 1: Inference with Proposed Regularizer.

Input: Initial estimate S^0 of S.
Output: Low-dimensional S and kernel representation C.
Parameters: Initial ρ^0 and maximum ρ_{max} penalty, with scale ρ_s.

- $S = S^0, \rho = \rho^0$;
while $\rho \leq \rho_{max}$ **do**
    while *not converged* **do**
        - Fix S and estimate C via closed-form solution of (7) using Algorithm 2;
        - Fix C and minimize (6) to update S using LM algorithm;
    - $\rho = \rho\rho_s$;

$C^T C$, which best explains the observation w_i. Here (6) is generally a non-convex problem, unless the Mercer-kernel is linear, and must therefore be solved using non-linear optimization techniques. In this work, we use the Levenberg-Marquardt algorithm for optimizing (6).

Notice that (6) only computes the pre-image for the feature space projections of the data points with which the non-linear manifold (matrix C) is learned. An extension to our formulation is desirable if one wants to use the learned non-linear manifold for denoising test data in a classic pre-image estimation framework. Although a valuable direction to pursue, it is out of scope of the present paper.

3.2 ROBUST DIMENSIONALITY REDUCTION

One can interpret sub-problem (7) as a robust form of KPCA where the kernel matrix has been corrupted with Gaussian noise and we want to generate its low-rank approximation. Although (7) is non-convex we can solve it in closed-form via singular value decomposition. This closed-form solution is outlined in Algorithm 2 and is based on the following theorem:

Theorem 1. *With* $\mathcal{S}^n \ni A \succeq 0$ *let* $A = U\Sigma U^T$ *denote its singular value decomposition. Then*

$$\min_L \ \frac{\rho}{2}||A - L^T L||^2_{\mathcal{F}} + \tau||L||_* \tag{8}$$

$$= \sum_{i=1}^n \left(\frac{\rho}{2}(\sigma_i - \gamma_i^{*2})^2 + \tau\gamma_i^* \right). \tag{9}$$

A minimizer L^* *of* (8) *is given by*

$$L^* = \Gamma^* U^T \tag{10}$$

with $\Gamma^* \in \mathcal{D}_+^n$, $\gamma_i^* \in \{\alpha \in \mathbb{R}_+ \mid p_{\sigma_i,\tau/2\rho}(\alpha) = 0\} \bigcup \{0\}$, *where* $p_{a,b}$ *denotes the depressed cubic* $p_{a,b}(x) = x^3 - ax + b$. $\mathcal{D}_+^n$ *is the set of n-by-n diagonal matrices with non-negative entries.*

Algorithm 2: Robust Dimensionality Reduction.

Input: Current estimate of S.
Output: Low-dimensional representation C.
Parameters: Current ρ and regularization strength τ.

- $[U \ \Lambda \ U^T]$ = Singular Value Decomposition of $K(S)$;
`// ` Λ ` is a diagonal matrix, storing ` N
` singular values ` λ_i ` of ` $K(S)$.
for $i = 1$ **to** N **do**
    - Find three solutions ($l_r : r \in \{1, 2, 3\}$) of:
$$l^3 - l\lambda_i + \frac{\tau}{2\rho} = 0 ;$$
    - set $l_4 = 0$;
    - $l_r = \max(l_r, 0) \ \forall r \in \{1, 2, 3, 4\}$;
    - $r = arg \min_r \{\frac{\rho}{2}||\lambda_i - l_r^2||^2 + \tau l_r\}$;
    - $\bar{\lambda}_i = l_r$;

- $C = \bar{\Lambda} U^T$;
`// ` $\bar{\Lambda}$ ` is diagonal matrix storing ` $\bar{\lambda}_i$.

Theorem 1 shows that each eigenvalue of the minimizer C^* of (7) can be obtained by solving a depressed cubic whose coefficients are determined by the corresponding eigenvalue of the kernel matrix and the regularization strength τ. The roots of each cubic, together with zero, comprise a set of candidates for the corresponding eigenvalue of C^*. The best one from this set is obtained by choosing the value which minimizes (9) (see Algorithm 2).

As elaborated in Section 2, problem (7) can be seen as regularizing sum of square root ($L_{1/2}$ norm) of the eigenvalues of the matrix $K(S)$. In a closely related work Zongben et al. (2012), authors advocate $L_{1/2}$ norm as a better approximation for the cardinality of a vector then the more commonly used L_1 norm. A closed form solution for $L_{1/2}$ regularization similar to our work was outlined in Zongben et al. (2012) and was shown to outperform the L_1 vector norm regularization for sparse coding. To that end, our Theorem 1 and the proposed closed form solution (Algo 2) for (7) can

Table 1: Performance comparison on missing data completion on Oil Flow Dataset: Row 1 shows the amount of missing data and subsequent rows show the mean and standard deviation of the error in recovered data matrix over 50 runs on 100 samples of oil flow dataset by: (1) The mean method (also the initialization of other methods) where the missing entries are replaced by the mean of the known values of the corresponding attributes, (2) 1-nearest neighbor method in which missing entries are filled by the values of the nearest point, (3) PPCA Tipping & Bishop (1999), (4) PKPCA of Sanguinetti & Lawrence (2006), (5)RKPCA Nguyen & De la Torre (2009) and our method.

p(del)	0.05	0.10	0.25	0.50
mean	13 ± 4	28 ± 4	70 ± 9	139 ± 7
1-NN	5 ± 3	14 ± 5	90 ± 20	NA
PPCA	$3.7 \pm .6$	9 ± 2	50 ± 10	140 ± 30
PKPCA	5 ± 1	12 ± 3	32 ± 6	100 ± 20
RKPCA	3.2 ± 1.9	8 ± 4	27 ± 8	83 ± 15
Ours	**2.3±2**	**6±3**	**22±7**	**70±11**

be seen as generalization of Zongben et al. (2012) to include the $L_{1/2}$ matrix norms for which a simplified proof is included in the Appendix A. It is important to note however, that the motivation and implication of using $L_{1/2}$ regularization in the context of non-linear dimensionality reduction are significantly different to that of Zongben et al. (2012) and related work Du et al. (2013); Zhao et al. (2014) which are designed for linear modeling of the causal factors. The core insight of using L_1 regularization in the feature space via the parametrization given in 3 facilitates a natural way for non-linear modeling of causal factors with low dimensionality while solving an inverse problem by making feature space tractable.

4 EXPERIMENTS

In this section we demonstrate the utility of the proposed algorithm. The aims of our experiments are twofold: (i) to compare our dimensionality reduction technique favorably with KPCA and its robust variants; and (ii) to demonstrate that the proposed non-linear dimensionality regularizer consistently outperforms its linear counterpart (a.k.a. nuclear norm) in solving inverse problems.

4.1 MATRIX COMPLETION

The nuclear norm has been introduced as a low rank prior originally for solving the matrix completion problem. Thus, it is natural to evaluate its non-linear extensions on the same task. Assuming $W \in R^{m \times n}$ to be the input matrix and Z a binary matrix specifying the availability of the observations in W, Algorithm 1 can be used for recovering a complete matrix S with the following choice of $f(W, Z, S)$:

$$f(W, Z, S) = \|Z \circ (W - S)\|_{\mathcal{F}}^2 \tag{11}$$

where $\circ$ represents Hadamard product.

To demonstrate the robustness of our algorithm for matrix completion problem, we choose 100 training samples from the oil flow dataset described in section 3.2 and randomly remove the elements from the data with varying range of probabilities to test the performance of the proposed algorithm against various baselines. Following the experimental setup as specified in Sanguinetti & Lawrence (2006), we repeat the experiments with 50 different samples of Z. We report the mean and standard deviation of the root mean square reconstruction error for our method with the choice of $\tau = 0.1$, alongside five different methods in Table 1. Our method significantly improves the performance of missing data completion compared to other robust extensions of KPCA Tipping & Bishop (1999); Sanguinetti & Lawrence (2006); Nguyen & De la Torre (2009), for every probability of missing data.

Although we restrict our experiments to least-squares cost functions, it is vital to restate here that our framework could trivially incorporate robust functions like the L_1 norm instead of the Frobenius norm — as a robust data term $f(W, Z, S)$ — to generalize algorithms like Robust PCA Wright et al. (2009) to their non-linear counterparts.

4.2 KERNEL NON-RIGID STRUCTURE FROM MOTION

Non-rigid structure from motion under orthography is an ill-posed problem where the goal is to estimate the camera locations and 3D structure of a deformable objects from a collection of 2D images which are labeled with landmark correspondences Bregler et al. (2000). Assuming $s_i(x_j) \in \mathbb{R}^3$ to be the 3D location of point x_j on the deformable object in the ith image, its orthographic projection $w_i(x_j) \in \mathbb{R}^2$ can be written as $w_i(x) = R_i s_i(x_j)$, where $R_i \in \mathbb{R}^{2 \times 3}$ is a orthographic projection matrix Bregler et al. (2000). Notice that as the object deforms, even with given camera poses, reconstructing the sequence by least-squares reprojection error minimization is an ill-posed problem. In their seminal work, Bregler et al. (2000) proposed to solve this problem with an additional assumption that the reconstructed shapes lie on a low-dimensional linear subspace and can be parameterized as linear combinations of a relatively low number of basis shapes. NRSfM was then cast as the low-rank factorization problem of estimating these basis shapes and corresponding coefficients.

Figure 2: Non-linear dimensionality regularisation improves NRSfM performance compared to its linear counterpart. Figure shows the ground truth 3D structures in red wire-frame overlaid with the structures estimated using: (a) proposed non-linear dimensionality regularizer shown in blue dots and (b) corresponding linear dimensionality regularizer (TNH) shown in black crosses, for sample frames of CMU mocap sequence. Red circles represent the 3D points for which the projections were known whereas squares annotated missing 2D observations. See text and Table 2 for details.

Recent work, like Dai et al. (2014); Garg et al. (2013a) have shown that the trace norm regularizer can be used as a convex envelope of the low-rank prior to robustly address ill-posed nature of the problem. A good solution to NRSfM can be achieved by optimizing:

$$\min_{S,R} \tau \|S\|_* + \sum_{i=1}^{F} \sum_{j=1}^{N} Z_i(x_j) \|w_i(x_j) - R_i s_i(x_j)\|_{\mathcal{F}}^2 \qquad (12)$$

where S is the shape matrix whose columns are $3N$ dimensional vectors storing the 3D coordinates $S_i(x_j)$ of the shapes and $Z_i(x_j)$ is a binary variable indicating if projection of point x_j is available in the image i.
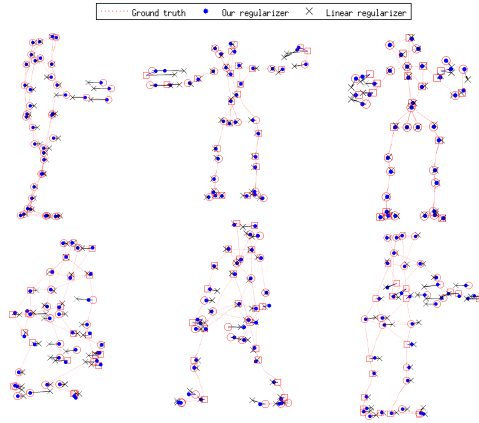
Assuming the projection matrices to be fixed, this problem is convex and can be exactly solved with standard convex optimization methods. Additionally, if the 2D projections $w_i(x_j)$ are noise free, optimizing (12) with very small τ corresponds to selecting the the solution — out of the many solutions — with (almost) zero projection error, which has minimum trace norm Dai et al. (2014). Thus henceforth, optimization of (12) is referred as the trace norm heuristics (TNH). We solve this problem with a first order primal-dual variant of the algorithm given in Garg et al. (2013a), which can handle missing data. The algorithm is detailed and compared favorably with the state of the art NRSfM approaches (based on linear dimensionality regularization) Appendix C.

A simple kernel extension of the above optimization problem is:

$$\min_{S,R} \tau \|\Phi(S)\|_* + \underbrace{\sum_{i=1}^{F} \sum_{j=1}^{N} Z_i(x_j) \|w_i(x_j) - R_i s_i(x_j)\|_{\mathcal{F}}^2}_{f(W,Z,R,S)} \qquad (13)$$

where $\Phi(S)$ is the non-linear mapping of S to the feature space using an RBF kernel.

With fixed projection matrices R, (13) is of the general form (2), for which the local optima can be found using Algorithm 1.

Table 2: 3D reconstruction errors for linear and non-linear dimensionality regularization with ground truth camera poses. Column 1 and 4 gives gives error for TNH while column (2-3) and (5-6) gives the corresponding error for proposed method with different width of RBF kernel. Row 5 reports the mean error over 4 sequences.

Dataset	No Missing Data			50% Missing Data		
	Linear	**Non-Linear**		**Linear**	**Non-Linear**	
		d_{max}	d_{med}		d_{max}	d_{med}
Drink	0.0227	0.0114	**0.0083**	0.0313	0.0248	**0.0229**
Pickup	0.0487	0.0312	**0.0279**	0.0936	0.0709	**0.0658**
Yoga	0.0344	**0.0257**	0.0276	0.0828	**0.0611**	0.0612
Stretch	0.0418	0.0286	**0.0271**	0.0911	**0.0694**	0.0705
Mean	0.0369	0.0242	**0.0227**	0.0747	0.0565	**0.0551**

4.2.1 RESULTS ON THE CMU DATASET

We use a sub-sampled version of CMU mocap dataset by selecting every 10th frame of the smoothly deforming human body consisting 41 mocap points used in Dai et al. (2014).[4]

In our experiments we use ground truth camera projection matrices to compare our algorithm against TNH. The advantage of this setup is that with ground-truth rotation and no noise, we can avoid the model selection (finding optimal regularization strength τ) by setting it low enough. We run the TNH with $\tau = 10^{-7}$ and use this reconstruction as initialization for Algorithm 1. For the proposed method, we set $\tau = 10^{-4}$ and use following RBF kernel width selection approach:

- Maximum distance criterion (d_{max}): we set the maximum distance in the feature space to be 3σ. Thus, the kernel matrix entry corresponding to the shape pairs obtained by TNH with maximum Euclidean distance becomes $e^{-9/2}$.
- Median distance criterion (d_{med}): the kernel matrix entry corresponding to the median euclidean distance is set to 0.5.

Following the standard protocol in Dai et al. (2014); Akhter et al. (2009), we quantify the reconstruction results with normalized mean 3D errors $e_{3D} = \frac{1}{\sigma F N} \sum_i \sum_j e_{ij}$, where e_{ij} is the euclidean distance of a reconstructed point j in frame i from the ground truth, σ is the mean of standard deviation for 3 coordinates for the ground truth 3D structures, and F, N are number of input images and number of points reconstructed.

Table 2 shows the results of the TNH and non-linear dimensionality regularization based methods using the experimental setup explained above, both without missing data and after randomly removing 50% of the image measurements. Our method consistently beats the TNH baseline and improves the mean reconstruction error by $\sim 40\%$ with full data and by $\sim 25\%$ when used with 50% missing data. Figure 2 shows qualitative comparison of the obtained 3D reconstruction using TNH and proposed non-lienar dimensionality regularization technique for some sample frames from various sequences. We refer readers to Appendix B for results with simultaneous reconstruction pose optimization.

5 CONCLUSION

In this paper we have introduced a novel non-linear dimensionality regularizer which can be incorporated into an energy minimization framework, while solving an inverse problem. The proposed algorithm for penalizing the rank of the data in the feature space has been shown to be robust to noise and missing observations. We have picked NRSfM as an application to substantiate our arguments and have shown that despite missing data and model noise (such as erroneous camera poses) our algorithm significantly outperforms state-of-the-art linear counterparts.

Although our algorithm currently uses slow solvers such as the penalty method and is not directly scalable to very large problems like dense non-rigid reconstruction, we are actively considering alternatives to overcome these limitations. An extension to estimate pre-images with a problem-

---

[4] Since our main goal is to validate the usefulness of the proposed non-linear dimensionality regularizer, we opt for a reduced size dataset for more rapid and flexible evaluation.

specific loss function is possible, and this will be useful for online inference with pre-learned low-dimensional manifolds.

Given the success of non-linear dimensionality reduction in modeling real data and overwhelming use of the linear dimensionality regularizers in solving real world problems, we expect that proposed non-linear dimensionality regularizer will be applicable to a wide variety of unsupervised inference problems: recommender systems; 3D reconstruction; denoising; shape prior based object segmentation; and tracking are all possible applications.

REFERENCES

Trine Julie Abrahamsen and Lars Kai Hansen. Input space regularization stabilizes pre-images for kernel pca de-noising. In *EEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2009.

Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Advances in neural information processing systems*, pp. 41–48, 2009.

Gokhan H Bakir, Jason Weston, and Bernhard Schölkopf. Learning to find pre-images. *Advances in neural information processing systems*, 16(7):449–456, 2004.

Christopher M Bishop and Gwilym D James. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 327(2):580–593, 1993.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999.

Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 690–696, 2000.

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.

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

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

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.

Amaury Dame, Victor Adrian Prisacariu, Carl Yuheng Ren, and Ian Reid. Dense reconstruction using 3d object shape priors. In *Computer Vision and Pattern Recognition*, pp. 1288–1295. IEEE, 2013.

Rong Du, Cailian Chen, Zhiyi Zhou, and Xinping Guan. L 1/2-based iterative matrix completion for data transmission in lossy environment. In *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on*, pp. 65–66. IEEE, 2013.

Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition*, pp. 1272–1279, 2013a.

Ravi Garg, Anastasios Roussos, and Lourdes Agapito. A variational approach to video registration with subspace constraints. *International journal of computer vision*, 104(3):286–314, 2013b.

Andreas Geiger, Raquel Urtasun, and Trevor Darrell. Rank priors for continuous non-linear dimensionality reduction. In *Computer Vision and Pattern Recognition*, pp. 880–887. IEEE, 2009.

Paulo FU Gotardo and Aleix M Martinez. Kernel non-rigid structure from motion. In *IEEE International Conference on Computer Vision*, pp. 802–809, 2011a.

Paulo FU Gotardo and Aleix M Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3065–3072. IEEE, 2011b.

Dong Huang, Ricardo Silveira Cabral, and Fernando De la Torre. Robust regression. In *European Conference on Computer Vision (ECCV)*, 2012.

Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

JT-Y Kwok and Ivor W Tsang. The pre-image problem in kernel methods. *IEEE Transactions on Neural Networks,*, 15(6):1517–1525, 2004.

Neil D Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *The Journal of Machine Learning Research*, 6:1783–1816, 2005.

Sebastian Mika, Bernhard Schölkopf, Alex J Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch. Kernel pca and de-noising in feature spaces. In *NIPS*, volume 4, pp. 7, 1998.

Minh Hoai Nguyen and Fernando De la Torre. Robust kernel principal component analysis. In *Advances in Neural Information Processing Systems*. 2009.

Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, New York, 2006.

Bryan Poling, Gilad Lerman, and Arthur Szlam. Better feature tracking through subspace constraints. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 3454–3461. IEEE, 2014.

Victor Adrian Prisacariu and Ian Reid. Nonlinear shape manifolds as shape priors in level set segmentation and tracking. In *Computer Vision and Pattern Recognition*, pp. 2185–2192. IEEE, 2011.

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

Ralph Tyrrell Rockafellar. *Conjugate duality and optimization*, volume 14. SIAM, 1974.

Guido Sanguinetti and Neil D Lawrence. Missing data in kernel pca. In *Machine Learning: ECML 2006*, pp. 751–758. Springer, 2006.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

John Wright, Arvind Ganesh, Shankar Rao, Yigang Peng, and Yi Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems*, pp. 2080–2088. 2009.

Qian Zhao, DeYu Meng, and ZongBen Xu. Robust sparse principal component analysis. *Science China Information Sciences*, 57(9):1–14, 2014.

Xiaowei Zhou, Can Yang, Hongyu Zhao, and Weichuan Yu. Low-rank modeling and its applications in image analysis. *ACM Computing Surveys (CSUR)*, 47(2):36, 2014.

Xu Zongben, Chang Xiangyu, Xu Fengmin, and Zhang Hai. L1/2 regularization: a thresholding representation theory and a fast solver. *IEEE Transactions on neural networks and learning systems*, 23(7):1013–1027, 2012.

A   PROOF OF THEOREM 3.1

Proof. We will prove theorem 1 by first establishing a lower bound for (8) and subsequently showing that this lower bound is obtained at L^* given by (10). The rotational invariance of the entering norms allows us to write (8) as:

$$\min_{\substack{\Gamma \in \mathcal{D}^n, \\ W^T W = I}} \frac{\rho}{2} ||\Sigma - W\Gamma^2 W^T||_{\mathcal{F}}^2 + \tau ||\Gamma||_*. \tag{14}$$

Expanding (14) we obtain

$$\frac{\rho}{2} \min_{\Gamma, W} \operatorname{tr}\left(\Sigma^2\right) - 2\operatorname{tr}\left(\Sigma W\Gamma^2 W^T\right) + \operatorname{tr}\left(\Gamma^4\right) + \frac{2\tau}{\rho}\sum_{i=1}^{n}\gamma_i \tag{15}$$

$$= \frac{\rho}{2} \min_{\Gamma, W} \sum_{i=1}^{n}\left(\sigma_i^2 + \gamma_i^4 + \frac{2\tau}{\rho}\gamma_i\right) - 2\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}^2\gamma_j^2\sigma_i \tag{16}$$

$$\geq \frac{\rho}{2} \min_{\Gamma} \sum_{i=1}^{n}\left(\sigma_i^2 - 2\gamma_i^2\sigma_i + \gamma_i^4 + \frac{2\tau}{\rho}\gamma_i\right) \tag{17}$$

$$= \frac{\rho}{2} \sum_{i=1}^{n} \min_{\gamma_i \geq 0}\left(\sigma_i^2 - 2\gamma_i^2\sigma_i + \gamma_i^4 + \frac{2\tau}{\rho}\gamma_i\right) \tag{18}$$

The inequality in (17) follows directly by applying Hölder's inequality to (16) and using the property that the column vectors w_i are unitary.

Next, with $L = \Gamma U^T$ in (8) we have

$$\frac{\rho}{2}||A - L^T L||_{\mathcal{F}}^2 + \tau ||L||_* = \frac{\rho}{2}||\Sigma - \Gamma^2||_{\mathcal{F}}^2 + \tau ||\Gamma||_*$$

$$= \sum_{i=1}^{n}\left(\frac{\rho}{2}(\sigma_i - \gamma_i^2)^2 + \tau\gamma_i\right). \tag{19}$$

Finally, since the subproblems in (18) are separable in γ_i, its minimizer must be KKT-points of the individual subproblems. As the constraints are simple non-negativity constraints, these KKT points are either (positive) stationary points of the objective functions or 0. It is simple to verify that the stationary points are given by the roots of the cubic function $p_{\sigma_i, \tau/2\rho}$. Hence it follows that there exists a γ_i^* such that

$$\frac{\rho}{2}\left(\sigma_i^2 - 2\gamma_i^2\sigma_i + \gamma_i^4 + \frac{2\tau}{\rho}\gamma_i\right) \geq \frac{\rho}{2}(\sigma_i - \gamma_i^{*2})^2 + \tau\gamma_i^*, \tag{20}$$

$\forall\gamma_i \geq 0$, which completes the proof. □

A.1   VALIDATING THE CLOSED FORM SOLUTION

Given the relaxations proposed in Section 2, our assertion that the novel trace regularization based non-linear dimensionality reduction is robust need to be substantiated. To that end, we evaluate our closed-form solution of Algorithm 2 on the standard oil flow dataset introduced in Bishop & James (1993).

This dataset comprises 1000 training and 1000 testing data samples, each of which is of 12 dimensions and categorized into one of three different classes. We add zero mean Gaussian noise with variance σ to the training data[5] and recover the low-dimensional manifold for this noisy training data S_σ with KPCA and contrast this with the results from Algorithm 2. An inverse width of the Gaussian kernel $\gamma = 0.075$ is used for all the experiments on the oil flow dataset.

It is important to note that in this experiment, we only estimate the principal components (and their variances) that explain the estimated non-linear manifold, i.e. matrix C by Algorithm 2, without reconstructing the denoised version of the corrupted data samples.

Both KPCA and our solution require model selection (choice of rank and τ respectively) which is beyond the scope of this paper. Here we resort to evaluate the performance of both methods under different parameters settings. To quantify the accuracy of the recovered manifold (C) we use following criteria:

---

[5]Note that our formulation assumes Gaussian noise in $K(S)$ where as for this evaluation we add noise to S directly.

Table 3: Robust dimensionality reduction accuracy by KPCA versus our closed-form solution on the full oil flow dataset. Columns from left to right represent: (1) standard deviation of the noise in training samples (2-3) Error in the estimated low-dimensional kernel matrix by (2) KPCA and (3) our closed-form solution, (4-5) Nearest neighbor classification error of test data using (4) KPCA and (5) our closed-form solution respectively.

STD	Manifold Error		Classification Error	
	KPCA	Our CFS	KPCA	Our CFS
.2	0.1099	**0.1068**	9.60%	9.60%
.3	0.2298	**0.2184**	19.90%	**15.70%**
.4	0.3522	**0.3339**	40.10%	**22.20%**
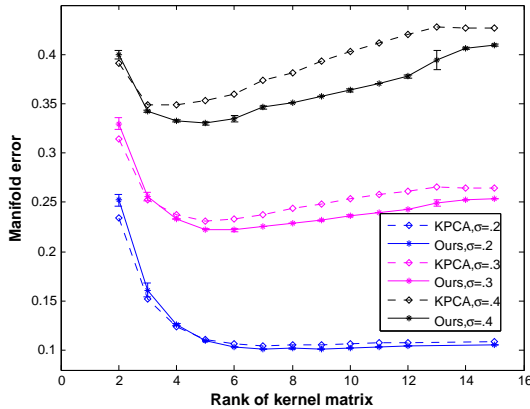
Figure 3: Performance comparison between KPCA and our Robust closed-form solution with dimensionality regularization on oil flow dataset with additive Gaussian noise of standard deviation σ. Plots show the normalized kernel matrix errors with different rank of the model. Kernel PCA results are shown in dotted line with diamond while ours are with solid line with a star. Bar-plot show the worst and the best errors obtained by our method for a single rank of recovered kernel matrix.

- Manifold Error : A good manifold should preserve maximum variance of the data — i.e. it should be able to generate a denoised version $K(S_{est}) = C^T C$ of the noisy kernel matrix $K(S_\sigma)$. We define the manifold estimation error as $\|K(S_{est}) - K(S_{GT})\|_\mathcal{F}^2$, where $K(S_{GT})$ is the kernel matrix derived using noise free data. Figure 3 shows the manifold estimation error for KPCA and our method for different rank and parameter τ respectively.[6]

- Classification error: The accuracy of a non-linear manifold is often also tested by the nearest neighbor classification accuracy. We select the estimated manifold which gives minimum Manifold Error for both the methods and report 1NN classification error (percentage of misclassified example) of the 1000 test points by projecting them onto estimated manifolds.

B KERNEL NRSFM WITH CAMERA POSE ESTIMATION

Extended from section 4.2

Table 4 shows the reconstruction performance on a more realistic experimental setup, with the modification that the camera projection matrices are initialized with rigid factorization and were refined with the shapes by optimizing (2). To solve NRSfM problem with unknown projection matrices, we parameterize each R_i with quaternions and alternate between refining the 3D shapes S and projection matrices R using LM. The regularization strength τ was selected for the TNH method by golden section search and parabolic interpolation for every test case independently. This ensures the best possible performance for the baseline. For our proposed approach τ was kept to 10^{-4} for all sequences for both missing data and full data NRSfM. This experimental protocol somewhat disadvantages the non-linear method, since its performance can be further improved by a judicious choice of the regularization strength.

---

[6]Errors from non-noisy kernel matrix can be replaced by cross validating the entries of the kernel matrix for model selection for more realistic experiment.

Table 4: 3D reconstruction errors for linear and non-linear dimensionality regularization with noisy camera pose initialization from rigid factorization and refined in alternation with shape. The format is same as Table 2.

Dataset	No Missing Data			50% Missing Data		
	Linear $\tau = \tau^*$	**Non-Linear** $\tau = 10^{-4}$		**Linear** $\tau = \tau^*$	**Non-Linear** $\tau = 10^{-4}$	
		d_{max}	d_{med}		d_{max}	d_{med}
Drink	0.0947	0.0926	**0.0906**	0.0957	0.0942	**0.0937**
Pickup	0.1282	0.1071	**0.1059**	0.1598	0.1354	**0.1339**
Yoga	0.2912	0.2683	**0.2639**	0.2821	**0.2455**	**0.2457**
Stretch	0.1094	0.1043	**0.1031**	**0.1398**	0.1459	0.1484
Mean	0.1559	0.1430	**0.1409**	0.1694	**0.1552**	**0.1554**

However our purpose is primarily to show that the non-linear method adds value even without time-consuming per-sequence tuning. To that end, note that despite large errors in the camera pose estimations by TNH and 50% missing measurements, the proposed method shows significant ($\sim 10\%$) improvements in terms of reconstruction errors, proving our broader claims that non-linear representations are better suited for modeling real data, and that our robust dimensionality regularizer can improve inference for ill-posed problems.

As suggested by Dai et al. (2014), robust camera pose initialization is beneficial for the structure estimation. We have used rigid factorization for initializing camera poses here but this can be trivially changed. We hope that further improvements can be made by choosing better kernel functions, with cross validation based model selection (value of τ) and with a more appropriate tuning of kernel width. Selecting a suitable kernel and its parameters is crucial for success of kernelized algorithms. It becomes more challenging when no training data is available. We hope to explore other kernel functions and parameter selection criteria in our future work.

We would also like to contrast our work with Gotardo & Martinez (2011a), which is the only work we are aware of where non-linear dimensionality reduction is attempted for NRSfM. While estimating the shapes lying on a two dimensional non-linear manifold, Gotardo & Martinez (2011a) additionally assumes smooth 3D trajectories (parametrized with a low frequency DCT basis) and a pre-defined hard linear rank constraint on 3D shapes. The method relies on sparse approximation of the kernel matrix as a proxy for dimensionality reduction. The reported results were hard to replicate under our experimental setup for a fair comparison due to non-smooth deformations. However, in contrast to Gotardo & Martinez (2011a), our algorithm is applicable in a more general setup, can be modified to incorporate smoothness priors and robust data terms but more importantly, is flexible to integrate with a wide range of energy minimization formulations leading to a larger applicability beyond NRSfM.

C   TNH ALGORITHM FOR NRSFM

In section 4.2, we have compared the proposed non-linear dimensionality reduction prior against a variant of Garg et al. (2013a) which handles missing data by optimizing:

$$\min_{S,R} \quad \tau\|S\|_* + \sum_{i=1}^{F}\sum_{j=1}^{N} Z_i(x_j)\|w_i(x_j) - R_i s_i(x_j)\|^2 \tag{21}$$

This problem is convex in S given noise free projection matrix R_i's but non-differentiable. To optimize (21), we first rewrite it in its primal-dual form by dualizing the trace norm [7] :

$$\max_{Q}\min_{S,R} \quad \tau <S,Q> + \sum_{i=1}^{F}\sum_{j=1}^{N} \quad Z_i(x_j)\|w_i(x_j) - R_i s_i(x_j)\|^2$$
$$s.t.\|Q\|_s \leq 1 \tag{22}$$

where $Q \in R^{\mathcal{X} \times N}$ stores the dual variables to S and $\|.\|_s$ represent spectral norm (highest eigenvalue) of a matrix.

---

[7]For more details on primal dual formulation and dual norm of the trace norm see Rockafellar (1974); Recht et al. (2010); Chambolle & Pock (2011).

Algorithm 3: Trace norm Heuristics.

Input: Initial estimates S^0, R^0 of S and R.
Output: Low-dimensional S and camera poses R.
Parameters: Regularization strength τ, measurements W and binary mask Z.

- $S = S^0$, $R = R^0$;

```
// set iteration count η step size σ and duals Q
```
- $\eta = 0$;
- $\sigma = 1/\tau$;
- $Q = 0$;

while *not converged* **do**

    ```// projection matrix estimation```
    - Fix S, Q and refine R_i for every image i with LM;

    ```// steepest descend update for S_ij for each point x_j and each frame i```
    for $i = 1$ **to** F **do**
        for $j = 1$ **to** N **do**
            - $S_{ij}^{\eta+1} = \left(I_{2\times 2} + \sigma(Z_{ij}R_i^T R_i)\right)^{-1} (S_{ij}^\eta - \sigma\tau Q_{ij}^\eta + \sigma R_i^T (Z_{ij} \circ w_{ij}))$;

    ```// accelerated steepest ascend update for Q```
    - $\bar{Q} = Q^\eta + \tau\sigma(2S^{n+1} - S^n)$;
    - UDV^T = singular value decomposition of Q^η;
    - $\bar{D} = \min(D, 1)$;
    - $Q^{\eta+1} = U\bar{D}V^T$;
    ```// Go to next iteration```
    - $\eta = \eta + 1$

Table 5: 3D reconstruction errors for different NRSfM approaches and our TNH Algorithm given ground truth camera projection matrices. Results for all the methods (except TNH) are taken from Dai et al. (2014).

Dataset	PTAAkhter et al. (2009)	CSF2Gotardo & Martinez (2011b)	BMMDai et al. (2014)	TNH
Drink	0.0229	**0.0215**	0.0238	0.0237
Pick-up	0.0992	0.0814	0.0497	**0.0482**
Yoga	0.0580	0.0371	0.0334	**0.0333**
Stretch	0.0822	0.0442	0.0456	**0.0431**

We choose quaternions to perametrize the 2×3 camera matrices R_i to satisfy orthonormality constraints as done in Garg et al. (2013a) and optimize the saddle point problem (22) using alternation. In particular, for a single iteration: (i) we optimize the camera poses R_i's using LM, (ii) take a steepest descend step for updating S and (ii) a steepest ascend step for updating Q which is followed by projecting its spectral norm to unit ball. Given ground truth camera matrices (without step (i)), alternation (ii-iii) can be shown to reach global minima of (22). Algorithm 3 outlines TNH algorithm.

As the main manuscript uses NRSfM only as a practical application of our non-linear dimensionality reduction prior, we have restricted our NRSfM experiments to only compare the proposed method against its linear counterpart. For the timely evaluation, the reported experiments we conducted on sub-sampled CMU mocap dataset. Here, we supplement the arguments presented in the main manuscript by favorably comparing the linear dimensionality reduction based NRSfM algorithm(TNH) to other NRSfM methods on full length CMU mocap sequences.