UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

CARNEGIE MELLON UNIVERSITY

Unifying Low-rank Models for Visual Learning

Ricardo da Silveira Cabral

Supervisors: Doctor Fernando De la Torre
Doctor João Paulo Salgado Arriscado Costeira
Co- Supervisor: Doctor Alexandre José Malheiro Bernardino

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury

Members of the Committee:

Doctor Fernando De la Torre
Doctor João Paulo Salgado Arriscado Costeira
Doctor Alexandre José Malheiro Bernardino
Doctor Aswin Sankaranarayanan
Doctor Andrew Fitzgibbon
Doctor Mário Alexandre Teles de Figueiredo

2015

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

CARNEGIE MELLON UNIVERSITY

Unifying Low-rank Models for Visual Learning

Ricardo da Silveira Cabral

Supervisors: Doctor Fernando De la Torre
Doctor João Paulo Salgado Arriscado Costeira
Co- Supervisor: Doctor Alexandre José Malheiro Bernardino

Thesis approved in public session to obtain the PhD Degree in

Electrical and Computer Engineering

Jury

Members of the Committee:

Doctor Fernando De la Torre
Doctor João Paulo Salgado Arriscado Costeira
Doctor Alexandre José Malheiro Bernardino
Doctor Aswin Sankaranarayanan
Doctor Andrew Fitzgibbon
Doctor Mário Alexandre Teles de Figueiredo

2015

Funding institution: Fundação para a Ciência e Tecnologia (SFRH/BD/33777/2009)

For Maria and Liz.

Abstract

Many problems in signal processing, machine learning and computer vision can be solved by learning low rank models from data. In computer vision, problems such as rigid structure from motion have been formulated as an optimization over subspaces with fixed rank. These *hard*-rank constraints have traditionally been imposed by a factorization that parameterizes subspaces as a product of two matrices of fixed rank. Whilst factorization approaches lead to efficient and kernelizable optimization algorithms, they have been shown to be NP-Hard in presence of missing data. Inspired by recent work in compressed sensing, hard-rank constraints have been replaced by *soft*-rank constraints, such as the nuclear norm regularizer. Vis-à-vis hard-rank approaches, soft-rank models are convex even in presence of missing data: but how is convex optimization solving a NP-Hard problem?

This thesis addresses this question by analyzing the relationship between hard and soft rank constraints in the unsupervised factorization with missing data problem. Moreover, we extend soft rank models to weakly supervised and fully supervised learning problems in computer vision. There are four main contributions of our work:

(1) The analysis of a new unified low-rank model for matrix factorization with missing data. Our model subsumes soft and hard-rank approaches and merges advantages from previous formulations, such as efficient algorithms and kernelization. It also provides justifications on the choice of algorithms and regions that guarantee convergence to global minima.

(2) A deterministic "rank continuation" strategy for the NP-hard unsupervised factorization with missing data problem, that is highly competitive with the state-of-the-art and often achieves globally optimal solutions. In preliminary work, we show that this optimization strategy is applicable to other NP-hard problems which are typically relaxed to convex semidefinite programs (e.g., MAX-CUT, quadratic assignment problem).

(3) A new soft-rank fully supervised robust regression model. This convex model is able to deal with noise, outliers and missing data in the input variables.

(4) A new soft-rank model for weakly supervised image classification and localization. Unlike existing multiple-instance approaches for this problem, our model is convex.

Keywords: Computer vision, Machine learning, Low-rank matrices, Convex optimization, Bilinear factorization, Augmented lagrange multiplier method, Image classification and localization, Weakly-supervised classification, Robust regression, Structure from motion.

Resumo

Vários problemas de processamento de sinal, aprendizagem automática e visão computacional podem ser resolvidos pela representação dos dados observados por matrizes *low-rank*. Particularmente na área da visão, problemas como *structure from motion* são formulados como problemas de optimização restrita a subespaços de rank fixo. Estas restrições são tradicionalmente impostas por uma factorização explícita dos subespaços, como um produto bilinear de duas matrizes. Embora desta imposição *estrita* das restrições de rank resultem algoritmos eficientes e com possibilidades de kernelização, os problemas de optimização resultantes desta modelação são NP-Hard quando sujeitos a omissões nos dados de entrada.

Motivados pelo progresso na área de *compressed sensing*, as restrições estritas têm vindo a ser substituídas por uma imposição *relaxada*, que adiciona à função de custo regularizadores como a norma nuclear. Em contraste com os modelos estritos, estes modelos são computacionalmente complexos, mas convexos mesmo na presença de dados parciais: mas como podem problemas NP-Hard ser resolvidos através da optimização convexa?

A presente tese visa abordar esta questão através da análise da relação entre os modelos estritos e relaxados, quando aplicados ao problema da factorização matricial com dados parciais. Adicionalmente, são propostos dois modelos relaxados para aprendizagem com supervisão fraca e total, aplicados ao problema da classificação de objectos e cenário em imagens. Os resultados podem ser sumarizados em quatro contribuições:

(1) A proposta e análise de um modelo unificado para o problema da factorização matricial low-rank com dados parciais. Este modelo engloba os modelos estritos e relaxados e apresenta vantagens de ambos, como a eficiência e a kernelização. O modelo permite ainda justificar os algoritmos utilizados e define regiões que garantem a convergência para mínimos globais.

(2) Uma estratégia determinística para optimização do problema da factorização com dados parciais, que demonstra resultados competitivos com as soluções existentes na literatura, alcançando o óptimo global com frequência. São ainda apresentados resultados preliminares que sugerem a aplicabilidade desta estratégia a problemas combinatóricos reformulados como programas semi-definidos positivos com restrições de rank (e.g., MAX-CUT, QAP).

(3) Um modelo relaxado de regressão e classificação robusta. Este modelo convexo é resiliente ao ruído, outliers e dados parciais nos dados de entrada.

(4) Um modelo relaxado para classificação e localização de objectos com supervisão fraca.

Palavras chave: Visão computacional, Aprendizagem automática, Matrizes de baixo rank, Optimização convexa, Factorização bilinear, Método ALM, Classificação e localização de imagens, Classificadores com supervisão fraca, Regressão robusta, Structure from motion.

Acknowledgments

During the last six years, as a direct consequence of having joined this PhD program, I have lived in 7 cities in 3 continents. So, it only seems fit that I get to write the acknowledgements part of this thesis several meters above ground, on a plane headed back to home, in the Azores islands. This is a fitting backdrop, because research means more often than not to "stand on the shoulders of giants" in the heights of their ideas, but also since these years have been as much of a voyage as any of these plane trips.

During this time, I had the incommensurable reward to get to know and work with so many inspiring people, from whom I've learned a great deal about the world, and ultimately about myself. To all of them, I dedicate the short paragraphs in this page in an admittedly disproportional way of saying thanks.

First and foremost, to Maria, the strongest person I have ever met. Your creativity, intelligence, brutal honesty and your incorruptible sense of ethics make you the most intriguing puzzle I've ever encountered, and after 11 years together it still feels I'm finding the edge pieces. This thesis is as much yours as it is mine.

Then, to my star team of advisors, without whom this would not have been possible. To Fernando, for all his "common sense" lessons. But more importantly, for the not-so-common sense, much more complex, life lessons on the inner workings of life as a researcher and dealing with people, the most complex machine of all. To Alex, for reading every single sentence I wrote this past 6 years, and writing as much back in reviews. Besides the fact that the reward to knowledge is knowledge itself, your work with robots reminded me many nights that computer vision actually matters to society and kept me motivated throughout these years. To JP, for being a visionnaire who thwarts adversities in the most creative ways, and for the friendship and *carinho* you have poured in the CMU-Portugal program and the people in it (me included). Your enthusiasm, as well as the way you look to the future and ultimately forge it yourself, are contagious and inspiring.

To my family. To Liz, for letting me be the 4-year old that I really am, and for showing me that I've already accomplished the most important thing of all. To my Mom, Dad and Sister, for being a constant beacon in my life whose love makes all research problems pale in comparison. To Ilda and João, for being as crazy as we are with moving and airplanes, and for making the cities we've lived in a home away from home. To Tia, Uncle Tony, Hugo, Fernando and Ginger, for being the closest family we've had geographically for several years, for their advice and for introducing us to the "American ways" and the delicious Thanksgivings!

To the special people that I've met while lurking around the hallways of ISR and the

Human Sensing lab, in no particular order: Ricardo Ferreira, Manuel Marques, Laura Trutoiu, Susana Brandão, Sabina Zejnilovic, Sérgio Pequito, João Mota, Ishan Misra, Xavier Perez-Sala, and Francisco Vicente. Thank you for sharing your passion of permutation and low-rank matrices, for many long nights of homeworks, but most importantly for friday "reading groups", great disposition, good and not-so-good moments and discussions, which had nothing to do with research whatsoever. I'm proud to call you friends.

Part of the research presented herein was shaped by collaboration with the awesomes Jayakorn Vongkulbhisal, Ehsan Adeli, Minh Hoai and Dong Huang. Professor João Xavier and Professor Yasutaka Furukawa, despite not being my advisors, graciously taught me many insights on convex optimization and 3D reconstruction. This thesis would not have been possible without Ana Mateus being a "superhero on a desk" ensuring all travel arrangements and paperwork are taken care of, so we can fully focus on research. I'm deeply appreciative of the support of CMU-Portugal (ICTI) program and Fundação para a Ciência (FCT), under the grant SFRH/BD/33777/2009. I would also like to acknowledge the committee members, Doctor Andrew Fitzgibbon, Professor Aswin Sankaranarayanan and Professor Mário Figueiredo, for their feedback.

"Computers are useless. They can only give you answers."

Pablo Picasso

Contents

Chapter 1

Introduction

1.1 The role of rank

The computer vision research area stands presently in an exciting time, with the ubiquity of imaging sensors in DSLRs, cellphones and laptops. Together with the advent of large computing power and global internet connectivity, these factors have eased the restrictions that limited amounts of data impose on the statistical learning of visual models, turning the so-called *curse of dimensionality* into a *blessing of dimensionality* [1]. Nonetheless, this new paradigm comes with its unique set of challenges: first, scalable algorithmic solutions are needed to harness this data, which cannot be stored or processed in a single computer; second, models that enforce Occam's razor's notion of simplicity become of utmost importance, so as to preserve model interpretability and avoid the risk of overfitting.

In this thesis, we study the topic of complexity penalization for visual learning tasks through rank minimization models. The use of rank criteria has been pervasive in computer vision applications as a mean of exploiting physical constraints of a model [2, 3, 4] or to minimize its complexity, be it in degrees of freedom [5] or in data redundancy [6, 7]. All these problems are directly or indirectly related to the problem of recovering a rank-k matrix

$\mathbf{Z}$ (see footnote[1] for notation) from a corrupted data matrix $\mathbf{X}$, by minimizing

$$\begin{aligned} \min_{\mathbf{Z}} \quad & f(\mathbf{X}-\mathbf{Z}) \\ \text{subject to} \quad & \text{rank}(\mathbf{Z}) = k, \end{aligned} \tag{1.1}$$

where $f(\cdot)$ denotes a loss function. Due to its intractability, the *hard*-rank constraint in (1.1) has typically been imposed by the inner dimensions of a bilinear factorization $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, as

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{X}-\mathbf{U}\mathbf{V}^\top). \tag{1.2}$$

The factorization approach in (1.2) has been popularized in computer vision by the seminal work on structure from motion of Tomasi and Kanade [2]. Since then, it has been applied to many problems, including non-rigid and articulated structure from motion, as well as photometric stereo [8] and motion segmentation [4], or even classification [9, 10, 11]. It has been shown that when the loss function $f(\cdot)$ is the Least-squares loss, *i.e.*, $f(\mathbf{X}-\mathbf{U}\mathbf{V}^\top) = \|\mathbf{X}-\mathbf{U}\mathbf{V}^\top\|_F^2$, then (1.2) does not have local minima and also that a closed form solution can be obtained via the Singular Value Decomposition (SVD) of $\mathbf{X}$ [12].

Unfortunately, this bilinear factorization approach has several caveats: The Least-squares loss is highly susceptible to outliers; also, the presence of missing data in $\mathbf{X}$ results in local minima. Outliers can be addressed with robust loss functions [7, 13] and optimal algorithms exist when missing data follows a Young diagram pattern [14]. However, missing data in computer vision typically exhibits random or band patterns, and factorization with missing data has been shown to be an NP-Hard problem [15], where many state-of-the-art

[1] Bold capital letters denote matrices (*e.g.*, $\mathbf{D}$). All non-bold letters denote scalar variables. d_{ij} denotes the scalar in the row i and column j of $\mathbf{D}$. $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ denotes the inner product between two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$. $||\mathbf{d}||_2^2 = \langle \mathbf{d}, \mathbf{d} \rangle = \sum_i d_i^2$ denotes the squared Euclidean Norm of the vector $\mathbf{d}$. $\text{tr}(\mathbf{A}) = \sum_i a_{ii}$ is the trace of $\mathbf{A}$. $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^\top\mathbf{A}) = \sum_{ij} a_{ij}^2$ designates the squared Frobenius Norm of $\mathbf{A}$. $||\mathbf{A}||_* = \sum_i \sigma_i$ designates the nuclear norm (sum of singular values σ_i) of $\mathbf{A}$. $\odot$ denotes the Hadamard or element-wise product. $\otimes$ denotes the Kronecker product. $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ denotes the identity matrix. $\text{diag}(\mathbf{X})$ is the vector of the diagonal elements of $\mathbf{X}$. $\text{Diag}(\mathbf{X})$ is a matrix containing only the diagonal elements of $\mathbf{X}$.

algorithms fail to even reach good local optima [16]. For this reason, the optimization of (1.2) remains an active research topic, with many works focusing on algorithms that are robust to initialization [3, 13, 17, 18, 19, 20], initialization strategies [21], or incorporating additional problem constraints to achieve better optima [8].

Recently, Candés and Recht [22] have stated that minimizing the rank function – under broad conditions of *incoherence*, *i.e.*, the unalignment of the singular vectors with the canonical axis – can be achieved by its convex surrogate, the nuclear norm. Initially proposed by Fazel [23], the nuclear norm permeated through many of the aforementioned computer vision problems such as structure from motion [17, 24, 25, 26], Robust PCA [6] and motion segmentation [27]. Here, the *soft*-rank regularization provided by the nuclear norm replaces the hard-rank constraints in the factorization approach of (1.2), by minimizing instead

$$\min_{\mathbf{Z}} \quad f(\mathbf{X} - \mathbf{Z}) + \lambda \|\mathbf{Z}\|_*, \tag{1.3}$$

where λ is a trade-off parameter between the error and the low-rank regularization induced by the nuclear norm $\|\mathbf{Z}\|_*$, the sum of singular values of $\mathbf{Z}$. We provide a simple intuition as to why the nuclear norm is in fact the largest possible convex underestimator of the rank function, as proved by [28]: Since the singular values of matrices are always positive, the nuclear norm can be interpreted as an ℓ_1-norm of the singular values. Under this interpretation, one can easily identify it as the convex envelope of the rank function, which is the cardinality (or ℓ_0-norm) of the singular values. To further understand why the singular value sparsity induced by the nuclear norm is important, let us consider completing the matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & ? \end{bmatrix}, \tag{1.4}$$

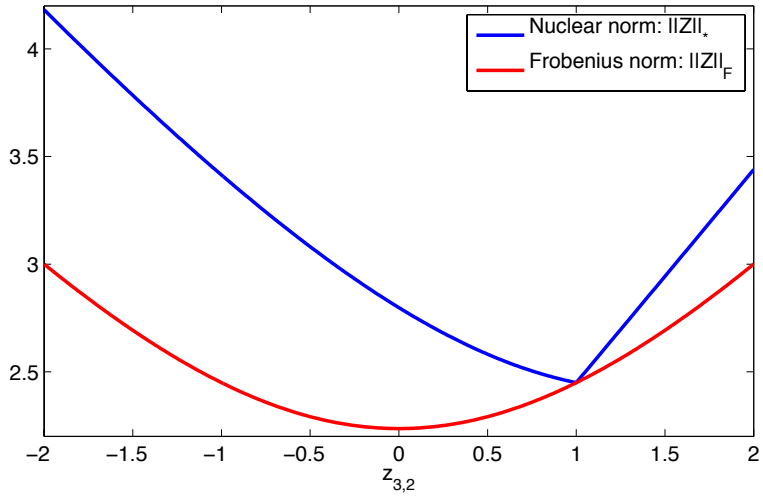


Figure 1.1: Comparison of Nuclear and Frobenius norms as function of one single unknown entry $z_{3,2}$ for the matrix in (1.4).

where only one entry $z_{3,2}$ is unknown such that the resulting rank is the smallest possible. The results shown in Figure 1.1 plot the nuclear norm and Frobenius norm of $\mathbf{Z}$ for all possible completions in a range around the value that minimizes its rank $z_{3,2} = 1$. In this case, the sparsity induced by the nuclear norm (ℓ_1-norm on the singular values) yields the optimal solution for $\mathbf{Z}$ with singular values $\sigma = [2.4495\ 0]$, a rank-1 matrix. In opposition, the Frobenius Norm (ℓ_2-norm of singular values) will set the entries to zero, thus leading to a solution with singular values $\sigma = [2.1358\ 0.6622]$, a rank-2 matrix. This key difference can be attributed to the fact that completing a matrix under the rank or nuclear norm favors the interaction between rows and columns to find a global solution, while the Frobenius norm treats each entry in the matrix independently (recall that $\|\mathbf{Z}\|_F^2 = \sum_{ij} z_{ij}^2$).

Contrary to hard-rank models, soft-rank regularization models have further extended the use of low-rank priors to many applications where the rank is not known *a priori*: colorization [29], subspace alignment [30] and clustering [31], segmentation [32], texture unwarping [33], camera calibration [34], tag refinement [35, 36], background modeling [6, 18] and tracking [37]. Soft-rank regularizers such as the nuclear norm or the max norm have also

been proposed in machine learning as good regularizers for classification [38]. Specifically, they has surfaced as a way to penalize complexity in image classification and regression tasks [11, 35, 39, 40, 41, 42, 43, 44], to reduce model degrees of freedom [45, 46, 47, 48], or to share properties among different classifiers [5, 47, 49].

Despite their convexity and theoretical results for the choice of λ [50], nuclear norm models such as the one in (1.3) also suffer from several drawbacks. On the one hand, it is unclear how to impose a certain rank in $\mathbf{Z}$: we showed in [51] that adjusting λ such that $\mathbf{Z}$ has a predetermined rank typically provides worse results than imposing this rank directly as in (1.2). Also, the inability to access the factorization of $\mathbf{Z}$ in (1.3) hinders the use of the "kernel trick" in classification and component analysis methods, and hence disallows for non-linear kernel extensions [52]. On the other hand, (1.3) is a Semidefinite Program (SDP). Current off-the-shelf SDP optimizers only scale to hundreds of variables, not amenable to the high dimensionality feature inputs typically found in computer vision problems. Several works [22, 50, 53, 54, 55, 56] ameliorate this issue by exploiting the fact that the proximal operator of the nuclear norm

$$\arg\min_{\mathbf{Z}} \quad \|\mathbf{X} - \mathbf{Z}\|_F^2 + \frac{1}{2}\|\mathbf{Z}\|_*, \tag{1.5}$$

has a closed form solution based on singular value thresholding. However, they still perform a SVD of $\mathbf{Z}$ in each iteration. Other approaches incrementally optimize (1.3) using gradient methods on the Grassmann manifold [57, 58, 59]. However, they rely on a rank selection heuristic, which fails when data is not missing at random. Zaid *et al.* [60] decompose the nuclear norm into a surrogate infinite-dimensional optimization, but their coordinate descent only applies to smooth losses $f(\cdot)$. Thus, nuclear norm approaches are currently unsuitable for handling dense, large scale datasets.

1.2 Main contributions

In summary, the main result of this thesis is that **soft rank nuclear norm models can be reformulated as hard rank factorization models** through the variational definition of the nuclear norm. While the variational definition was previously known in the literature [18, 61], we are the first to propose a unification of soft and hard-rank approaches in computer vision under one formulation.

Several implications stem from the unification of soft and hard-rank models, as we are able to analyze the conditions under which both approaches are equivalent: for soft models, we bring advantages such as scalability and kernelization, and show their limitations on problems where the output rank is predetermined. For hard models, we propose a deterministic "rank continuation" strategy for the NP-Hard factorization problem that avoids local optima in a significant number of cases when the rank is known *a priori*. We extend this strategy to the case of Binary Quadratic Problems such as the Quadratic Assignment Problem, which can be reformulated as rank-1 problems.

Additionally, we propose new soft-rank models for visual weakly supervised learning and fully supervised learning, two settings with different levels of information in the training data they are provided.

First, we study the problem of weakly supervised multi-label image classification, where images have been labeled with several present classes but their location in the image is not known. For this problem, we propose a convex matrix completion model specifically tailored to visual data. We also provide two alternative algorithms for optimizing this model as well as their convergence proofs. Our model can easily cope with labeling errors and missing data, background noise and partial occlusions. Moreover, it allows for learning latent individual representations for all classes in the dataset. Thus, we can recover localization information without the need for fully supervised training data with localization information.

Experimental validation on several datasets shows that our method outperforms state-of-the-art classification algorithms, while effectively capturing each class appearance.

Second, a fully supervised robust regression model, which learns a direct association from data to labels. For this case, we develop the theory of Robust Regression (RR). This framework applies to a variety of problems in computer vision including robust linear discriminant analysis, regression with missing data, and multi-label classification. Our framework is both convex and able to explicitly deal with missing data and outliers in the data. These advantages are contrary to existing discriminative methods, which fail to account for outliers that are common in realistic training sets due to occlusion, specular reflections or noise. Several synthetic and real examples with applications to head pose estimation from images, image and video classification and facial attribute classification with missing data are used to illustrate the benefits of RR.

1.3 Organization

The remainder of this thesis is organized as follows.

- In Chapter 2, we show that nuclear norm (soft rank) formulations can be reformulated as factorization (hard rank) formulations and unify them in a single model. We propose an augmented lagrange multiplier algorithm to solve the unified model and show that this equivalence result makes the kernelization of some nuclear norm models trivial. We then split the use of soft and hard-rank models into two regions of our unified model: when rank is known a priori or when rank is known to be low but not precisely known. This work has been published in [51].
- In Chapter 3, we propose two new soft-rank models for visual learning tasks such as object classification, detection, by exploiting the fact that data is known to be low-rank but it's specific rank is not known beforehand. This chapter contains work published in [44, 62, 63, 64].
- In Chapter 4 we focus on problems where rank is predetermined or known *a priori*. We show the limitations of soft rank models on these problems, and present "rank continuation", a deterministic strategy that empirically attains good solutions in the problems of factorization and graph matching. This chapter contains work published in [51] and extended in [65].
- Our conclusions and possible directions for future work are presented in Chapter 5. There, we also restate our major contributions and discuss their current limitations.

Chapter 2

Relation between soft and hard-rank constraints in low rank models

2.1 Definition and unification of soft and hard-rank models

As mentioned in Chapter 1, finding models that favor low rank solutions is an essential tool for solving computer vision and machine learning problems: low rank representations allow for reducing degrees of freedom, exploiting redundancy, and enforcing simplicity when representating shape, appearance or motion. There are two main approaches for imposing low-rank, which we will formally define as hard-rank and soft-rank models.

Definition 1 (Hard-rank models)**.** Optimization models that aim to recover a rank-k matrix $\mathbf{Z} \in \mathbb{R}^{M \times N}$ from a data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ according an error function $f(\cdot)$, as

$$\begin{aligned} \min_{\mathbf{Z}} \quad & f(\mathbf{X} - \mathbf{Z}) \\ \text{subject to} \quad & \text{rank}(\mathbf{Z}) = k. \end{aligned} \tag{2.1}$$

This constraint is typically directly imposed on the solution by optimizing a bilinear product

$\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ and specifying the inner dimensions of this product as k, as

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{X} - \mathbf{U}\mathbf{V}^\top). \tag{2.2}$$

Definition 2 (Soft-rank models)**.** Optimization models that aim to recover a rank-k matrix $\mathbf{Z} \in \mathbb{R}^{M\times N}$ from a data matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ according an error function $f(\cdot)$, and a low-rank solution is sought. Thus, the problem is regularized by adding to the cost function a regularizer such as the nuclear norm, as

$$\min_{\mathbf{Z}} \quad f(\mathbf{X} - \mathbf{Z}) + \lambda\|\mathbf{Z}\|_*. \tag{2.3}$$

In this chapter, we show that nuclear norm (soft rank) formulations can be reformulated as factorization (hard rank) formulations and thus unify them in a single model. Let us start by considering the nuclear norm problem in (2.3) with convex $f(\cdot)$: without loss of generality, we can rewrite (2.3) as the SDP [66]

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{B},\mathbf{C}} \quad & f(\mathbf{X} - \mathbf{Z}) + \frac{\lambda}{2}\left(\mathrm{tr}(\mathbf{B}) + \mathrm{tr}(\mathbf{C})\right) \\ \text{subject to} \quad & \mathbf{Q} = \begin{bmatrix} \mathbf{B} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{C} \end{bmatrix} \succeq 0. \end{aligned} \tag{2.4}$$

For any positive semidefinite matrix $\mathbf{Q}$, we can write $\mathbf{Q} = \mathbf{R}\mathbf{R}^\top$ for some $\mathbf{R}$. Thus, we can

replace matrix $\mathbf{Q}$ in (2.4) by

$$\mathbf{Q} = \begin{bmatrix} \mathbf{B} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{U}^\top & \mathbf{V}^\top \end{bmatrix}, \tag{2.5}$$

where $\mathbf{U} \in \mathbb{R}^{M \times r}$, $\mathbf{V} \in \mathbb{R}^{N \times r}$ and $r \leq \min(N, M)$ upper bounds $\operatorname{rank}(\mathbf{Z})$. Merging (2.5) into (2.4) yields

Definition 3 (Unified model)**.**

$$\min_{\mathbf{U},\mathbf{V}} \quad f(\mathbf{X} - \mathbf{U}\mathbf{V}^\top) + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right), \tag{2.6}$$

where the SDP constraint was dropped because it is satisfied by construction. This reformulation seems counterintuitive, as we changed the convex problem in (2.3) into a non-convex one, which may be prone to local minima (*e.g.*, in the case of missing data under the least-squares loss [15]). However, we show that the existence of local minima in (2.6) depends only on the dimension r imposed on matrices $\mathbf{U}$ and $\mathbf{V}$. We extend the analysis of Burer and Monteiro [67] to prove that:

Theorem 1. *Let $f(\mathbf{X}-\mathbf{Z})$ be convex in $\mathbf{Z}$ and $\mathbf{Z}^*$ be an optimal solution of the convex nuclear norm model in (2.3) for a given λ and let $\operatorname{rank}(\mathbf{Z}^*) = k^*$. Then, any solution $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ of (2.6) with $r \geq k^*$ is a global minima solution of (2.3).*

Theorem 1 (which we prove in Appendix A) immediately allows us to draw one conclusion: By application of the variational property of the nuclear norm [66],

$$\|\mathbf{Z}\|_* \quad = \min_{\mathbf{Z}=\mathbf{U}\mathbf{V}^\top} \frac{1}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right), \tag{2.7}$$

many soft-rank models can be reformulated into hard-rank models. That is, the factorization

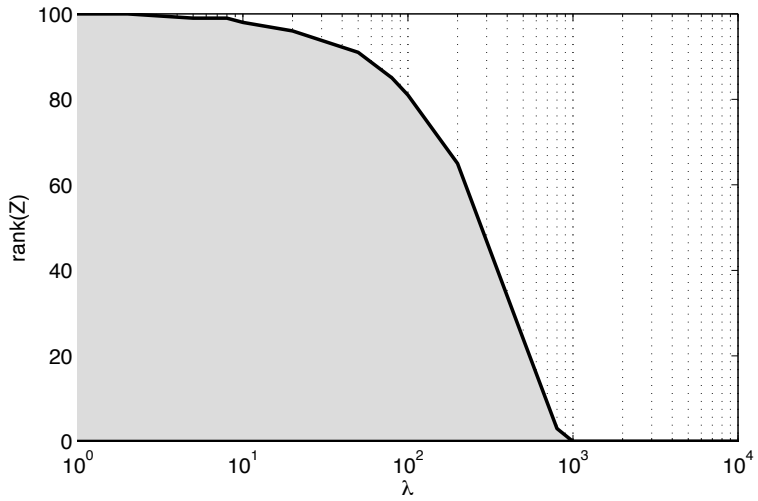


Figure 2.1: Region of equivalence between factorization (2.6) and nuclear norm approaches (2.3) for a 100×100 random matrix and least-squares loss. When factorization is initialized in the white area, it is equivalent to the result obtained with the nuclear norm (black line). When the rank is known *a priori*, directly imposing $r = k$ in the factorization approach of (2.6) (gray area) is less prone to local minima than the unregularized problem (2.2) and provides better results than selecting λ in the nuclear norm model (2.3) such that the output rank is k.

and the nuclear norm models in (2.2) and (2.3) are special cases of (2.6). Fig. 2.1 illustrates the result of Theorem 1 in a synthetic case. We plot the output rank of $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ in (2.6) as a function of λ for a random 100×100 matrix $\mathbf{X}$ with all entries sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, 1)$, no missing data and $f(\cdot)$ is the least squares loss $\|\mathbf{X} - \mathbf{Z}\|_F^2$: the factorization approach in (2.2) corresponds to the case where $\lambda = 0$ and r is fixed, whilst the nuclear norm in (2.3) outputs an arbitrary rank k^* as a function of λ (the black curve). According to Theorem 1, for any $r \geq k^*$ (white area), optimizing (2.6) is equivalent to (2.3). On the other hand, when $r < k^*$ (grey area), the conditions of Theorem 1 are no longer valid and thus (2.6) can be prone to local minima.

A special case of Theorem 1 has been used to recommend the use of nuclear norm approaches in the machine learning community by Mazumder *et al.* [61]. However, their analysis is restricted to the least-squares loss and the case where the rank is not known *a priori* (*i.e.*, white area of Fig. 2.1). Our analysis instead extends to other convex loss

functions and is motivated by the observation that many computer vision problems live in the grey area of Fig. 2.1. That is, their output rank k is predetermined by a domain-specific constraint (*e.g.*, in Structure from Motion $k = 4$ [2]).

The visual interpretation of Theorem 1 in Fig. 2.1 shows two clear regions of operation of our unified model. As such, for the remainder of this thesis, we will consider the use of soft and hard-rank models as two separate regions of our unified model: when rank is known a priori or when rank is known to be low but not precisely known. We advocate the use of our unified model in (2.6) for both cases over the typical soft and hard-rank models, based on two arguments:

When the output rank is unconstrained (white area of Fig. 2.1) soft-rank models should be used, but we can always choose $r \geq k^*$ such that (2.6) provides equivalent results to (2.3). Using the result in Theorem 1 and the analysis of Burer and Monteiro [67], we propose an ALM algorithm in Section 2.2 using the unified model in (2.6) that has the scalability advantages of factorization approaches, yet it is guaranteed to attain the global optima of the original nuclear norm model. Also, we show that this equivalence result makes the kernelization of some nuclear norm models trivial in Section 2.3. In Chapter 3, we propose several new soft-rank models for visual learning tasks and show our unified model is faster than state-of-the-art algorithms for optimizing nuclear norm models.

When the output rank is known *a priori* (gray area of Fig. 2.1) hard-rank models should be used, but optimizing (2.6) is preferable to (2.2) and (2.3). As we will show in Chapter 4, optimizing (2.6) is less prone to local minima than the unregularized problem (2.2). On the other hand, selecting λ in the nuclear norm model (2.3) such that the output rank k is the desired value typically leads to worse results than directly imposing $r = k$ in (2.6). Based on this analysis, we propose in Sec. 4.1 a "rank continuation" strategy, and empirically show it is able to attain global optimality in several scenarios.

2.2 An ALM algorithm for the unified model

Nuclear norm models have extended the use of low-rank priors to many applications where $\mathbf{Z}$ is low rank but its exact value is not known *a priori* [34, 44, 50]. In this section, we propose an algorithm for solving (2.6) and show that its complexity is lower than proximal methods [53] for optimizing the nuclear norm model in (1.3). For the remainder of this section, we focus our attention in the LS loss

$$f(\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})) = \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 = \sum_{ij} (w_{ij}(x_{ij} - z_{ij}))^2, \tag{2.8}$$

and the L1 loss

$$f(\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})) = \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_1 = \sum_{ij} |w_{ij}(x_{ij} - z_{ij})|, \tag{2.9}$$

where $\mathbf{W} \in \mathbb{R}^{M \times N}$ is a positive weight matrix that can be used to denote missing data (*i.e.*, $w_{ij} = 0$). We note, however, that the results in Theorem 1 also apply to many other losses such as the Huber [13, 24] and hinge loss [11, 42].

One important factor to take into account when optimizing (2.6) for the LS and L1 losses is that when either $\mathbf{U}$ or $\mathbf{V}$ is fixed, the remaining part of (2.6) becomes convex, even in presence of a missing data pattern specified by $\mathbf{W}$. However, it has been reported that pure alternation approaches for this problem are prone to flatlining [3, 16, 19]. For smooth losses such as the LS, this can be circumvented by performing gradient steps jointly in $\mathbf{U}, \mathbf{V}$ [19]. Alternatively, we propose an Augmented Lagrange Multiplier (ALM) method for two reasons: 1) Theorem 1 and the analysis in [67] can be used to prove our ALM's convergence to global optima of (1.3) when $r \geq k^*$, and 2) its applicability to the non-smooth L1 norm. Let us

rewrite (2.6) as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{U},\mathbf{V}} \quad & f(\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})) + \frac{\lambda}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right) \\ \text{subject to} \quad & \mathbf{Z} = \mathbf{U}\mathbf{V}^\top, \end{aligned} \tag{2.10}$$

and its corresponding augmented lagrangian as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{U},\mathbf{V},\mathbf{Y}} \quad & f(\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})) + \frac{\lambda}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right) \\ & + \langle \mathbf{Y}, \mathbf{Z} - \mathbf{U}\mathbf{V}^\top \rangle + \frac{\rho}{2} \|\mathbf{Z} - \mathbf{U}\mathbf{V}^\top\|_F^2, \end{aligned} \tag{2.11}$$

where $\mathbf{Y}$ are Lagrange multipliers and ρ is a penalty parameter to improve convergence [53]. This method exploits the fact that the solution for each subproblem in $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ can be efficiently solved in closed form. For $\mathbf{U}$ and $\mathbf{V}$, the solution is obtained by equating the derivatives of (2.11) in $\mathbf{U}$ and $\mathbf{V}$ to $\mathbf{0}$. For known $\mathbf{U}$ and $\mathbf{V}$, $\mathbf{Z}$ can be updated by solving

$$\min_{\mathbf{Z}} \quad f(\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})) + \frac{\rho}{2} \|\mathbf{Z} - \left(\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y} \right)\|_F^2, \tag{2.12}$$

which can be done in closed form by the element-wise shrinkage operator $\mathcal{S}_\mu(x) = \max(0, x - \mu)$, as

$$\begin{aligned} \mathbf{Z} = \mathbf{W} \odot & \left(\mathbf{X} - \mathcal{S}_{\rho^{-1}}(\mathbf{X} - \mathbf{U}\mathbf{V}^\top + \rho^{-1}\mathbf{Y}) \right) \\ & + \overline{\mathbf{W}} \odot (\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y}), \end{aligned} \tag{2.13}$$

for the L1 loss, or

$$\begin{aligned} \mathbf{Z} = \mathbf{W} \odot & \left(\frac{1}{2+\rho} \left(2\mathbf{X} + \rho(\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y}) \right) \right) \\ & + \overline{\mathbf{W}} \odot (\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y}), \end{aligned} \tag{2.14}$$

for the LS loss. Here, $\overline{w}_{ij} = 1, \forall_{ij} w_{ij} \neq 0$ and 0 otherwise. The resulting algorithm is summarized in Alg. 1 and its full derivation is presented in Sec. 2.2.1. Contrary to pure alternated methods, our numerical experiments show that this method is not prone to flatlining due to the joint optimization being gradually enforced by the lagrange multipliers $\mathbf{Y}$.

Algorithm 1 ALM method for optimizing (2.6)

Input: $\mathbf{X}, \mathbf{W} \in \mathbb{R}^{M \times N}$, params μ, λ, initialization of ρ

  while not converged **do**

    while not converged **do**

      Update $\mathbf{U} = (\rho\mathbf{Z} + \mathbf{Y})\,\mathbf{V}\left(\rho\mathbf{V}^\top\mathbf{V} + \lambda\mathbf{I}_r\right)^{-1}$

      Update $\mathbf{V} = (\rho\mathbf{Z} + \mathbf{Y})^\top\,\mathbf{U}\left(\rho\mathbf{U}^\top\mathbf{U} + \lambda\mathbf{I}_r\right)^{-1}$

      Update $\mathbf{Z}$ via (2.13) for L1 loss or (2.14) for LS loss

    end while

    $\mathbf{Y} = \mathbf{Y} + \rho(\mathbf{Z} - \mathbf{U}\mathbf{V}^\top)$

    $\rho = \min\left(\rho\mu, 10^{20}\right)$

  end while

Output: Complete Matrix $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$

Assuming without loss of generality that $\mathbf{X} \in \mathbb{R}^{M \times N}$ and $M > N$, we have that exact state-of-the-art methods for SVD (*e.g.*, Lanczos bidiagonalization algorithm with partial reorthogonalization) take a flop count of $O(MN^2 + N^3)$. The most computational costly step in our ALM method are the matrix multiplications in the update of $\mathbf{U}$ and $\mathbf{V}$, which take $O(MNr + Nr^2)$ if done naively. Given that typically $k^* \leq r \ll \min(M, N)$ and k^* can be efficiently estimated [68], Alg. 1 provides significant computational cost savings when compared to proximal methods which use SVDs [53].

We note that there are several recent results in optimization which tackle the complexity issue of SVDs in proximal methods for the nuclear norm. For instance, there has been recent work on online methods for factorization [69], as well as randomized or incremental SVDs [70]. Also, when using a projected sub-gradient method one can easily avoid the cost of SVD using a polar decomposition of the variable Z which can be obtained by Halley's method [66]. If the singular values are away from zero, this is much faster than the original SVD algorithm. Also, there are approaches that minimize models for RPCA in linear time. For instance, [71] solve an initial smaller problem of the dimension of the rank r and then calculate the remainder of the matrix using projections based on the calculated singular vector estimates. However, our result is still relevant in this case for solving the initial

problem, as the rank r may still be a large number even if considerably smaller than the matrix dimensions $\min(M, N)$. Moreover, our result allows very scalable solutions recently obtained for factorization methods (e.g., [72]) to be applied to nuclear norm models by resorting to its variational definition.

2.2.1 Full derivation of Algorithm 1

We provide the full derivation of Alg. 1 in this section. Let us start by transforming the problem

$$\min_{\mathbf{U},\mathbf{V}} \quad \|\mathbf{W} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^\top)\|_1 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right), \tag{2.15}$$

into the equivalent problem

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{U},\mathbf{V}} \quad & \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_1 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) \\ \text{subject to} \quad & \mathbf{Z} = \mathbf{U}\mathbf{V}^\top, \end{aligned} \tag{2.16}$$

and write its augmented lagrangian function, as

$$\begin{aligned} \min_{\mathbf{Z},\mathbf{U},\mathbf{V}} \mathcal{L} = \quad & \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_1 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right) + \\ & \langle \mathbf{Y}, \mathbf{Z} - \mathbf{U}\mathbf{V}^\top \rangle + \frac{\rho}{2}\|\mathbf{Z} - \mathbf{U}\mathbf{V}^\top\|_F^2, \end{aligned} \tag{2.17}$$

where $\mathbf{Y} \in \mathbb{R}^{d \times n}$ is a Lagrange multiplier matrix, and ρ is a penalty parameter [53]. We solve (2.17) by an iterative Gauss-Siedel method on $\mathbf{Z}, \mathbf{U}, \mathbf{V}$, solved by the subproblems

$$\mathbf{Z}^{(k+1)} = \arg\min_{\mathbf{Z}} \mathcal{L}(\mathbf{Z}^{(k)}, \mathbf{U}, \mathbf{V}, \mathbf{Y}, \rho) \tag{2.18}$$

$$\mathbf{U}^{(k+1)} \arg\min_{\mathbf{U}} \mathcal{L}(\mathbf{Z}, \mathbf{U}^{(k)}, \mathbf{V}, \mathbf{Y}, \rho), \tag{2.19}$$

$$\mathbf{V}^{(k+1)} = \arg\min_{\mathbf{V}} \mathcal{L}(\mathbf{Z}, \mathbf{U}, \mathbf{V}^{(k)}, \mathbf{Y}, \rho), \tag{2.20}$$

where k is the index of iterations. At iteration $k = 0$, the entries of variables $\mathbf{U}, \mathbf{V}, \mathbf{Z}$ are initialized i.i.d. from a standard normal distribution and $\mathbf{Y}, \rho$ are initialized as

$$\mathbf{Y}^{(0)} = \mathbf{0} \tag{2.21}$$

$$\rho^{(0)} = 10^{-5} \tag{2.22}$$

After initialization,(2.18)-(2.20) are solved sequentially until convergence. In the following subsections, we will derive the solutions of each of these subproblems.

After each Gauss-Siedel convergence, the Lagrange Multiplier matrix $\mathbf{Y}$ is updated by a gradient ascent step

$$\mathbf{Y}^{(k+1)} = \mathbf{Y}^{(k)} + \rho^{(k)}(\mathbf{Z} - \mathbf{U}\mathbf{V}^\top), \tag{2.23}$$

where the penalty variable ρ is updated by the expression

$$\rho^{(k+1)} = \mu\rho^{(k)}, \tag{2.24}$$

and $\mu > 1$ is a constant. A larger μ imposes stronger enforcement of the constraint $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, therefore faster convergence of the outer loop, but may result in poor performance of the inner Gauss-Siedel loop and vise versa. In our experiments, we chose $\mu = 1.05$.

Solving for U Fixing $\mathbf{Z}$ and $\mathbf{V}$, the subproblem (2.19) is reduced to the problem

$$\mathcal{L}(\mathbf{U}) \propto \frac{\lambda}{2}\|\mathbf{U}\|_F^2 + \langle \mathbf{Y}, \mathbf{Z} - \mathbf{U}\mathbf{V}^\top \rangle + \frac{\rho}{2}\|\mathbf{Z} - \mathbf{U}\mathbf{V}^\top\|_F^2, \tag{2.25}$$

whose closed-form solution can be obtained by equating the derivative of (2.25) to $\mathbf{0}$, resulting in

$$\mathbf{U} = (\rho\mathbf{Z} + \mathbf{Y})\,\mathbf{V}\left(\rho\mathbf{V}^\top\mathbf{V} + \lambda\mathbf{I}_r\right)^{-1} \tag{2.26}$$

Solving for V Fixing $\mathbf{Z}$ and $\mathbf{U}$, the subproblem (2.20) is reduced to the problem

$$\mathcal{L}(\mathbf{V}) \propto \frac{\lambda}{2}\|\mathbf{V}\|_F^2 + \langle \mathbf{Y}, \mathbf{Z} - \mathbf{U}\mathbf{V}^\top \rangle + \frac{\rho}{2}\|\mathbf{Z} - \mathbf{U}\mathbf{V}^\top\|_F^2, \tag{2.27}$$

whose closed-form solution can be obtained by equating the derivative of (2.27) to $\mathbf{0}$, resulting in

$$\mathbf{V} = (\rho\mathbf{Z} + \mathbf{Y})^\top \mathbf{U} \left(\rho\mathbf{U}^\top\mathbf{U} + \lambda\mathbf{I}_r\right)^{-1} \tag{2.28}$$

Solving for Z Fixing $\mathbf{U}$, $\mathbf{V}$, the cost function of subproblem (2.18) can be rewritten in an equivalent problem

$$\min_{\mathbf{Z}} \quad \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_1 + \frac{\rho}{2}\|\mathbf{Z} - \left(\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y}\right)\|_F^2, \tag{2.29}$$

which can be done in closed form from the fact that $\mathbf{0}$ is in the expression of the subdifferential of (2.29). Using the element-wise shrinkage operator $\mathcal{S}_\mu(x) = \max(0, x - \mu)$, this condition can be written as

$$\begin{aligned} \mathbf{Z} = \mathbf{W} \odot \big(\mathbf{X} - \mathcal{S}_{\rho^{-1}}(\mathbf{X} - \mathbf{U}\mathbf{V}^\top + \rho^{-1}\mathbf{Y})\big) \\ + \overline{\mathbf{W}} \odot (\mathbf{U}\mathbf{V}^\top - \rho^{-1}\mathbf{Y}), \end{aligned} \tag{2.30}$$

where $\overline{w}_{ij} = 1, \forall_{ij} w_{ij} \neq 0$, and 0 otherwise.

Stopping criteria For the outer loop in Alg. 1 in the main paper, the iteration is not terminated until the equality constraint $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$ is satisfied up to a given tolerance. In our experiments, we used $\|\mathbf{Z} - \mathbf{U}\mathbf{V}^\top\|_F \leq 10^{-9}\|\mathbf{W} \odot \mathbf{M}\|_F$. For the inner loop, since the global optimum solution is found for (2.18)-(2.20), the objective function monotonically decreases. As such, in our experiments the stopping criteria for the inner Gauss-Siedel loop combines two items:

1. Small decrease of $\mathcal{L}(\cdot)$: $\frac{\|\mathcal{L}(\cdot)^{(k)} - \mathcal{L}(\cdot)^{(k-1)}\|}{\|\mathcal{L}(\cdot)^{(k-1)}\|} \leq 10^{-10}$;
2. Maximum number of iterations is reached: 5000.

2.3 Kernelization of nuclear norm methods

An important implication of Theorem 1 of Section 2 is that in (2.6) we can solve (1.3) with access to $\mathbf{U}, \mathbf{V}$. This has one immediate implication: we can exploit the kernel trick in nuclear norm models by resorting to their equivalent factorized formulation, which makes kernel extensions straightforward. For example, we show in Sec. 2.3 an example extension for Robust LDA [44].

Kernelization of PCA

When using our unified model (2.6), the kernelization of PCA follows trivially from the classical technique, by replacing covariance matrices by their kernel versions. Moreover, it is interesting to note that the regularization terms imposed by the nuclear norm in (2.6) correspond to the terms $\lambda \mathbf{I}_r$ in the solutions for $\mathbf{U}, \mathbf{V}$ in Alg. 1. This gives interpretability to the "trick" of adding such terms in component analysis techniques to ensure proper conditioning of the inverse when the approximated covariance matrices $\mathbf{U}^\top \mathbf{U}$ and $\mathbf{V}^\top \mathbf{V}$ are singular due to the small sample size problem [52], as it can be interpreted as a soft rank regularization being applied to the component analysis model.

Kernelization for RLDA

In this section, we illustrate kernelization for the case of Robust LDA. Let $\mathbf{X} \in \mathbb{R}^{D \times N}$ be a matrix where each column is a vectorized data sample from one of C classes. D denotes the number of features and N the number of samples. $\mathbf{G} \in \mathbb{R}^{N \times C}$ is an indicator matrix such that $g_{ij} = 1$ if $\mathbf{x}_i$ belongs to class j, and 0 otherwise. LDA can then be formulated as [52],

$$\min_{\mathbf{U},\mathbf{V}} \quad \|(\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}}(\mathbf{G} - \mathbf{U}\mathbf{V}^\top \mathbf{X})\|_F^2 \tag{2.31}$$

Note that in this case of an L1 robust function such as in [44], selecting the rank in $\mathbf{U}, \mathbf{V}$ will not yield an eigen-problem as the LS loss case, but instead a problem which may contain

several local minima. Therefore, it is best to use the soft regularization instead, as

$$\min_{\mathbf{Z}} \quad \|(\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}}(\mathbf{G} - \mathbf{Z}\mathbf{X})\|_1 + \lambda \|\mathbf{Z}\|_* \tag{2.32}$$

The analysis provided in our paper allows us to reformulate (2.32) as

$$\min_{\mathbf{U},\mathbf{V}} \quad \|(\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}}(\mathbf{G} - \mathbf{U}\mathbf{V}^\top \mathbf{X})\|_1 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right), \tag{2.33}$$

Having access to factors $\mathbf{U}, \mathbf{V}$ and using the Mercer theorem [73], we can express $\mathbf{V}$ as a linear combination of the data, as

$$\mathbf{V} = \mathbf{X}\alpha, \tag{2.34}$$

which yields

$$\begin{aligned} \min_{\mathbf{U},\alpha} \quad & \|(\mathbf{G}^\top \mathbf{G})^{-\frac{1}{2}}(\mathbf{G} - \mathbf{U}\alpha^\top \mathbf{X}^\top \mathbf{X})\|_1 + \\ & \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \mathrm{tr}(\alpha^\top \mathbf{X}^\top \mathbf{X}\alpha)\right), \end{aligned} \tag{2.35}$$

Since $\mathbf{X}$ always appears in a dot product $\mathbf{X}^\top \mathbf{X}$ in (2.35), we can easily replace it by a kernel matrix, thus allowing for non-linear extensions of this method.

Chapter 3

Using soft-rank models when rank is not known *a priori*

As seen in Chapter 1, soft-rank models have extended the use of low-rank priors to many applications where the rank is not known *a priori*. In particular, soft-rank regularizers such as the nuclear norm or the max norm have been proposed in machine learning as good regularizers for classification [38]. These have surfaced as a way to penalize complexity in image classification and regression tasks [11, 35, 39, 40, 41, 43], to reduce model degrees of freedom [45, 46, 47, 48], for recovering localization cues from classification [39, 42], or to share properties among different classifiers [5, 47, 49]. In this chapter, we describe our model contributions to these visual learning tasks, split by their level of supervision. To summarize, the main contributions of this chapter are threefold:

- In Sec. 3.1, we propose a new nuclear norm model for fully supervised robust regression, which learns a direct association from data to labels. This convex framework applies to a variety of problems in computer vision including robust linear discriminant analysis, regression with missing data, and multi-label classification. Several synthetic and real examples with applications to head pose estimation from images, image and video classification and facial attribute classification with missing data are used to illustrate

the benefits of RR. This work was published in [44] and currently under review in a journal submission.

- In Sec. 3.2, we propose a new nuclear norm model for weakly supervised image classification, where images have been labeled with several present classes but their location in the image is not known. We cast the problem under a matrix completion transduction model, which is able to classify and localize images. Unlike existing discriminative methods, our model is convex and is robust to labeling errors, background noise and partial occlusions. We propose a Fixed-Point Continuation (FPC) algorithm for solving matrix completion and prove its the convergence in Appendix B. FPC algorithms for matrix completion had been proposed previously in the literature, but their convergence had not been proven to extend to constrained problems. Experimental validation on several datasets shows that our method outperforms state-of-the-art classification algorithms, while effectively capturing each class appearance. This work was originally published in [62] and extended in [63].
- In Sec. 3.3, we analyze the application of the unified model proposed in Chapter 2 to unsupervised learning tasks such as background subtraction, using a Robust PCA model. Experiments show that the ALM method proposed in Chapter 2 is both faster and more accurate than state-of-the-art nuclear norm algorithms for this task.

3.1 Fully supervised learning as a robust regression problem

Linear and non-linear regression models have been applied to solve a number of computer vision problems (*e.g.*, classification [74], pose estimation [75]). Although widely used, a major drawback of existing regression approaches is their lack of robustness to outliers and noise, which are common in realistic training sets due to occlusion, specular reflections or image noise. To better understand the lack of robustness, we consider the problem of learning a

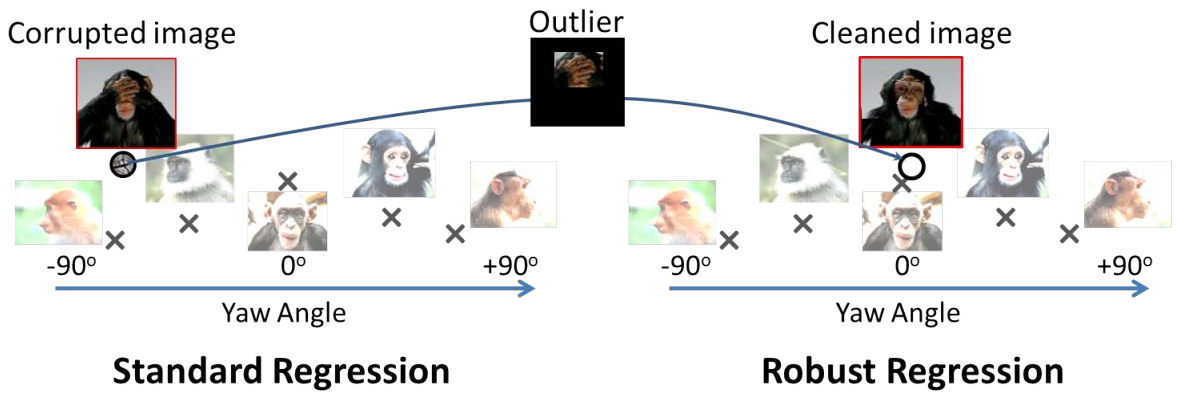


Figure 3.1: Predicting the yaw angle of the monkey head from image features. Note the image features (image pixels) contain outliers (hands of the monkey). (Left) Standard regression: projects a partially occluded frontal face image *directly* onto the head pose subspace and fails to estimate the correct yaw angle; (Right) Robust regression removes the intra-sample outlier and projects only the cleaned input image without biasing the yaw angle estimation.

linear regressor from image features $\mathbf{X}$ to pose angles $\mathbf{Y}$ (see Fig. 3.1) by minimizing

$$\min_{\mathbf{T}} \|\mathbf{Y} - \mathbf{T}\mathbf{X}\|_F^2. \tag{3.1}$$

In the training stage, we learn the mapping $\mathbf{T}$, and in testing we estimate the pose by projecting the features $\mathbf{x}_{te}$ of the test image, $\mathbf{T}\mathbf{x}_{te}$.

Standard regression, Eq. (3.1), is optimal under the assumption that the error, $\mathbf{E} = \mathbf{Y} - \mathbf{T}\mathbf{X}$, is normally distributed. The Least Squares (LS) estimate is the most efficient unbiased estimate of $\mathbf{T}$ in presence of Gaussian noise. This is the well known Gauss-Markov theorem [76]. However, a small number of gross outliers can arbitrarily bias the estimate of the model's parameters ($\mathbf{T}$). It is important to notice that in training and testing $\mathbf{X}$ is assumed to be noise free. However, a single outlier in either training or testing can bias the projection because LS projects the data *directly* onto the subspace of $\mathbf{T}$. The dot product of $\mathbf{x}_{te}$ with each row of $\mathbf{T}$ (*i.e.*, $\mathbf{T}\mathbf{x}_{te}$) can be largely biased by only a single outlier. For this reason, existing discriminative methods lack robustness to outliers.

The problem of robustness in regression has been studied thoroughly in statistics, and the last decades have witnessed a fast paced development of the so-called robust methods (*e.g.*, [77, 78, 79]). For instance, M-estimators [77] assume the error has a heavy tail and typically re-weight the whole sample inversely proportional to it using different influence functions. That is, some robust approaches minimize a weighted regression $\sum_{i=1}^{n} w_i \|\mathbf{y}_i - \mathbf{T}\mathbf{x}_i\|_2^2$, where w_i weighs the whole sample. Other robust approaches replace the sum (or the mean) by a more robust measure such as the median (*e.g.*, least median of squares) [80] or trimmed mean (*e.g.*, least trimmed square) [78]. However, all of the aforementioned traditional robust approaches for regression differ from the problem addressed in this chapter in two ways: (1) these approaches do not model the error in $\mathbf{X}$ but in $\mathbf{Y} - \mathbf{T}\mathbf{X}$, (2) they mostly consider sample-outliers (*i.e.*, the whole image is an outlier). This work proposes an intra-sample robust regression (RR) method that explicitly accounts for outliers in $\mathbf{X}$. Our work is related to errors in variables (EIV) models (*e.g.*, [81, 82, 83]). However, unlike existing EIV models, RR does not require a prior estimate of the noise and all parameters are automatically estimated.

In addition to reducing the influence of noise and outliers in regression, we extend RR to be able to deal with missing data in regression, wherein some elements of $\mathbf{X}$ are unknown. This is a common issue in computer vision applications, since unknown elements typically correspond to unobserved local image features. Surprisingly, this problem has been relatively unexplored in the computer vision literature. We illustrate the power of RR in several computer vision tasks including head pose estimation from images, facial attribute detection with missing data and robust LDA for multi-label image classification.

3.1.1 Related work

Extensive literature exists on robust methods for regression. Huber [77] introduced M-estimation for regression, providing robustness to sample outliers. Rousseeuw and Leroy

proposed Least Trimmed Squares [78], which explicitly finds a data subset that minimizes the squared residual sum. Parallel to developments in the statistics community, the idea of subset selection has also flourished in many computer vision applications. Consensus approaches such as RANSAC [84] (and its Maximum Likelihood (ML) and M-estimator variants [85, 86]) randomly subsample input data to construct a tentative model. Model parameters are updated when a new configuration produces smaller inlier error than its predecessors. In spite of accurate parameter estimates, even in the presence of several outliers, these methods heavily rely on the assumption that model generation from a data subset is computationally inexpensive and inlier detection can be done adequately. Moreover, the aforementioned methods do not tackle *intra-sample* outliers, *i.e.*, partial sample corruptions.

To deal with noise in the variables, Error-In-Variable (EIV) approaches have been proposed, see [82] for an overview. However, existing EIV approaches rely on strong parametric assumptions for the errors. For instance, orthogonal regression assumes that the variance of errors in the input and response variables are identical [87] or their ratio is known [88]. Under these assumptions, orthogonal regression can minimize the Gaussian error orthogonal to the learned regression vectors. Grouping-based methods [89] assume that errors are respectively i.i.d. among the input and response variables, so that one can split the data into groups and suppress the errors by computing differences of the group sum, geometric means or instrument variables. Moment-based methods [90] learn the regression by estimating high-order statistics, *i.e.*, moments, from i.i.d. data. Likelihood-based methods [83] learn a reliable regression when the input and response variables follow a joint, normal and identical distribution. Total Least Square (TLS) [82] and its nonlinear generalization [91] solve for additive/multiple terms that enforce the correlation between the input and response variables. TLS-based methods relax the assumption in previous methods to allow correlated and non-identically distributed errors. Nevertheless, they still rely on parametric assumptions on the error. Unfortunately, in typical computer vision applications, errors caused by

occlusion, shadow and edges seldom fit such distributions.

Although regression and classification are single-handled by our framework, several authors have addressed solely the issue of robust classification. The majority of these methods can be cast as robust extensions of Fisher/Linear Discriminant Analysis (FDA/LDA), where the empirical estimation of the class mean vectors and covariance matrices are replaced by their robust counterparts such as MVE estimators [92], MCD estimators [93] and S-estimators [94, 95]. In machine learning, several authors [96, 97] have proposed a worst-case FDA/LDA by minimizing the upper bound of the LDA cost function to increase the separation ability between classes under unbalanced sampling. As in previous work on robust regression, these methods are only robust to sample-outliers.

Our work is more related to recent work in computer vision. Fidler and Leonardis [98] robustify LDA for intra-sample outliers. In the training stage, [98] computed PCA on the training data, replaced the minor PCA components by a robustly estimated basis, and combined the two basis into a new one. Then, the data was projected onto the combined basis and LDA is computed. During testing, [98] first estimates the coefficients of a test data on the recombined basis by sub-sampling the data elements using [99]. Finally, the class label of the test data is determined by applying learned LDA on the estimated coefficients. Although outliers outside of the PCA subspace can be suppressed, [98] does not address the problem of learning LDA with outliers in the PCA subspace of the training data. Zhu and Martinez [100] proposed learning a SVM with missing data and robust to outliers. In [100], the possible values for missing elements are modeled by a Gaussian distribution, and such that for each class, the input data with all possible missing elements spans an affine subspace. The decision plane of the robustified SVM jointly maximizes the between-class margin while minimizing the angle between the decision plane and the class-wise affine subspaces. However, [100] requires the location of the outliers to be known. In contrast to previous works, our RR enjoys several advantages: (1) it is a convex approach; (2) no assumptions, aside from

sparsity, are imposed on the outliers, which makes our method general; (3) it automatically cleans the intra-sample outliers in the training data while learning a classifier.

3.1.2 Robust Regression (RR)

Let $\mathbf{X} \in \mathbb{R}^{d_x \times n}$ be a matrix containing n d_x-dimensional samples possibly corrupted by outliers. Formally, $\mathbf{X} = \mathbf{D} + \mathbf{E}$, where $\mathbf{D} \in \mathbb{R}^{d_x \times n}$ is matrix containing the underlying noise-free component and $\mathbf{E} \in \mathbb{R}^{d_x \times n}$ models all noise including outliers. In regression problems, one learns a mapping $\mathbf{T}$ from $\mathbf{X}$ to an output $\mathbf{Y} \in \mathbb{R}^{d_y \times n}$. The outliers or the noise-free component $\mathbf{D}$ are unknown, so existing methods use $\mathbf{X}$ in the estimation of $\mathbf{T}$. In presence of outliers, this results in a biased estimation of $\mathbf{T}$. Our RR solves this problem by explicitly decomposing $\mathbf{X}$ into $\mathbf{D} + \mathbf{E}$, and only computing $\mathbf{T}$ using the clean free data $\mathbf{D}$. RR solves the following optimization problem

$$\begin{aligned} \min_{\mathbf{T},\mathbf{D},\mathbf{E}} \quad & \frac{\eta}{2} \left\| \mathbf{W}(\mathbf{Y} - \mathbf{T}\hat{\mathbf{D}}) \right\|_F^2 + rank(\mathbf{D}) + \lambda \|\mathbf{E}\|_0 \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \quad \hat{\mathbf{D}} = [\mathbf{D}; \mathbf{1}^T], \end{aligned} \tag{3.2}$$

where $\mathbf{W} \in \mathbb{R}^{d_y \times d_y}$ is a diagonal matrix that weights the output dimensions, $\mathbf{T} \in \mathbb{R}^{d_y \times (d_x+1)}$ is the regression matrix (the extra dimension is for the regression bias term). η and λ are scalars that weight the first and third term in Eq. (3.2) respectively. RR explicitly avoids projecting the outlier matrix $\mathbf{E}$ to the output space by learning the regression $\mathbf{T}$ only from the augmented noise-free data $\hat{\mathbf{D}} = [\mathbf{D}; \mathbf{1}^T] \in \mathbb{R}^{(d_x+1) \times n}$. Observe that there are infinite possible decompositions of $\mathbf{X}$ into $\mathbf{D}$ and $\mathbf{E}$, so RR adds the second and third terms in Eq. (3.2) to constrain the possible solutions. The second term constrains $\mathbf{D}$ to lie in a low-dimensional subspace, which is a good prior for naturally occurring data [52]. The third term encourages $\mathbf{E}$ to be sparse.

It is important to notice that RR is different from cleaning the data using RPCA and then computing LS-regression on the clean data, because RR *cleans* the input data $\mathbf{X} = \mathbf{D} + \mathbf{E}$ in

a supervised manner; that is, the data $\mathbf{D}$ will preserve the subspace of $\mathbf{X}$ that is maximally correlated with $\mathbf{Y}$. For this reason, the outlier component $\mathbf{E}$ computed by RR is able to correct outliers both inside and outside the subspace spanned by $\mathbf{D}$ (see the experiment in section 3.1.6).

The original form of RR, Eq. (3.2), is cumbersome to solve because the rank and cardinality operators are non-convex, so these operators are respectively relaxed to their convex surrogates: the nuclear norm and the ℓ_1-norm. Using this relaxation Eq. (3.2) is rewritten as

Definition 4 (Robust regression model)**.**

$$\begin{aligned} \min_{\mathbf{T},\mathbf{D},\mathbf{E}} \quad & \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{D}\|_* + \lambda\|\mathbf{E}\|_1 \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \hat{\mathbf{D}} = [\mathbf{D}; \mathbf{1}^T]. \end{aligned} \tag{3.3}$$

This problem can be efficiently optimized using an Augmented Lagrange Muliplier (ALM) technique, wherein Eq. (3.3) is rewritten as

$$\begin{aligned} \min_{\mathbf{T},\mathbf{D},\hat{\mathbf{D}},\mathbf{E}} \quad & \frac{\eta}{2}\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{D}\|_* + \lambda\|\mathbf{E}\|_1 \\ & + \langle \Gamma_1, \mathbf{X}-\mathbf{D}-\mathbf{E}\rangle + \frac{\mu_1}{2}\|\mathbf{X}-\mathbf{D}-\mathbf{E}\|_F^2 \\ & + \langle \Gamma_2, \hat{\mathbf{D}} - [\mathbf{D}; \mathbf{1}^T]\rangle + \frac{\mu_2}{2}\|\hat{\mathbf{D}} - [\mathbf{D}; \mathbf{1}^T]\|_F^2, \end{aligned} \tag{3.4}$$

where $\Gamma_1 \in \mathbb{R}^{d_x \times n}$ and $\Gamma_2 \in \mathbb{R}^{(d_x+1)\times n}$ are Lagrange multiplier matrices, and μ_1 and μ_2 are the penalty parameters. For each of the four matrices $\{\mathbf{T}, \mathbf{D}, \hat{\mathbf{D}}, \mathbf{E}\}$ to be solved in Eq. (3.4), the cost function is convex if the remainder three matrices are kept fixed. Details of the ALM method to minimize Eq. (3.4) are given in Alg. 2.

3.1.3 Robust LDA: extending RR for classification

Classification problems can be cast as a particular case of binary regression, where each sample in $\mathbf{X}$ belongs to one of c classes. The goal is then to learn a mapping from $\mathbf{X}$ to labels indicating the class membership of the data points. LDA learns a linear transformation that maximizes inter-class separation while minimizing intra-class variance, and typical solutions are based on solving a generalized eigenvalue problem. However, when learning from high-dimensional data such as images ($n < d_x$), LDA typically suffers from the small sample size problem. While there are several approaches to solve the small sample size problem (*e.g.*, regularization), a more fundamental solution is to relate the LDA problem to a reduced-rank LS problem [101]. LS-LDA [101] directly maps $\mathbf{X}$ to the class labels by minimizing

$$\min_{\mathbf{T}} \left\| (\mathbf{Y}\mathbf{Y}^T)^{-1/2}(\mathbf{Y} - \mathbf{T}\mathbf{X}) \right\|_F^2, \tag{3.5}$$

where $\mathbf{Y} \in \mathbb{R}^{c \times n}$ is a binary indicator matrix, such that $y_{ij} = 1$ if $\mathbf{x}_j$ belongs to class i, otherwise $y_{ij} = 0$. The normalization factor $\mathbf{W} = (\mathbf{Y}\mathbf{Y}^T)^{-1/2}$ compensates for different number of samples per class. $\mathbf{T} \in \mathbb{R}^{c \times d_x}$ is a reduced rank regression matrix, which typically has rank $c - 1$ (if the data is centered). After $\mathbf{T}$ is learned, a test datum $\mathbf{x}_{te} \in \mathbb{R}^{d_x \times 1}$ is projected by $\mathbf{T}$ onto the c dimensional output space spanned by $\mathbf{T}$, then the class label of the test data $\mathbf{x}_{te}$ is assigned to its maximum value or using k-NN if one wishes an additional degree of nonlinearity.

When $\mathbf{X}$ is corrupted by outliers, Eq. (3.5) suffers from the same bias problem as standard regression. RR, Eq. (3.3), can be directly applied to Eq. (3.5), yielding

Definition 5 (Robust LDA model)**.**

$$\begin{aligned} \min_{\mathbf{T},\mathbf{D},\mathbf{E}} \quad & \frac{\eta}{2}\left\|(\mathbf{Y}\mathbf{Y}^T)^{-1/2}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2 + \|\mathbf{D}\|_* + \lambda\|\mathbf{E}\|_1 \\ s.t. \quad & \mathbf{X} = \mathbf{D} + \mathbf{E}, \quad \hat{\mathbf{D}} = [\mathbf{D}; \mathbf{1}^T]. \end{aligned} \tag{3.6}$$

This Robust LDA formulation can be easily solved as a special case of RR (Alg. 2).

3.1.4 Robustness in testing data

In the previous sections, we have assumed that the training set was corrupted by outliers and noise. Similarly, the test data might contain outliers, and as in the case of training, RR removes outliers before projection. Let us refer to $\mathbf{X}_{te} \in \mathbb{R}^{d_x \times n_{te}}$ as a set of test samples (n_{te} samples), and $\mathbf{Y}_{te} \in \mathbb{R}^{d_y \times n_{te}}$ the estimated label, the subscript te denote the test data. Observe that this is a non-trivial problem because the test label matrix $\mathbf{Y}_{te}$ is not available to provide the supervised information.

Consider Eq. (3.3) without the first supervised term,

$$\begin{aligned} \min_{\mathbf{D}_{te},\mathbf{E}_{te}} \quad & \|\mathbf{D}_{te}\|_* + \lambda \|\mathbf{E}_{te}\|_1 \\ s.t. \quad & \mathbf{X}_{te} = \mathbf{D}_{te} + \mathbf{E}_{te}, \end{aligned} \tag{3.7}$$

where $\mathbf{D}_{te} \in \mathbb{R}^{d_x \times n_{te}}$ is the cleaned test data, $\mathbf{E}_{te} \in \mathbb{R}^{d_x \times n_{te}}$ is the noise/outlier matrix, and λ is the positive scalar determined in training (see Eq. (3.3)).

Eq. (3.7) is convex and equivalent to the nuclear norm RPCA model in [50]. However, RPCA is an unsupervised technique and it can only clean the outliers/noise that are orthogonal to $\mathbf{X}_{te}$. We will refer to this noise as out-of-subspace noise. If we are interested in removing the error within the subspace of $\mathbf{X}_{te}$, this can be done by using the cleaned training data $\mathbf{D}$. In the training stage, $\mathbf{D}$ is optimized to have maximum correlation with

Algorithm 2 ALM algorithm for solving RR Eq. (3.3)

Input: $\mathbf{X}$, $\mathbf{Y}$, parameters η (a positive scalar weights term $\left\|\mathbf{W}(\mathbf{Y}-\mathbf{T}\hat{\mathbf{D}})\right\|_F^2$), λ (a positive scalar weights term $\|\mathbf{E}\|_1$), ρ (a positive scalar for updating the Lagrange coefficients), γ (a positive scalar for regularizing the solution to $\mathbf{T}$).

Initialization:$\mathbf{D}^{(0)}=\mathbf{X}$, $\hat{\mathbf{D}}^{(0)}=[\mathbf{D}^{(0)};\mathbf{1}^T]$, $\mathbf{E}^{(0)}=\mathbf{X}-\mathbf{D}^{(0)}$, $\mathbf{T}^{(0)}=(\hat{\mathbf{D}}^{(0)}(\hat{\mathbf{D}}^{(0)})^{\mathbf{T}}+\gamma\mathbf{I}_{d_x+1})^{-1}\mathbf{Y}(\hat{\mathbf{D}}^{(0)})^T$;

Lagrange Multiplier Initialization: $\Gamma_1^{(0)}=\frac{\mathbf{X}}{\|\mathbf{X}\|_2}$,$\Gamma_2^{(0)}=\frac{\mathbf{D}^{(0)}}{\|\mathbf{D}^{(0)}\|_2}$,$\mu_1^{(0)}=\frac{dn}{4}\|\mathbf{X}\|_1$, $\mu_2^{(0)}=\frac{dn}{4}\|\mathbf{D}^{(0)}\|_1$.

while $\frac{\left\|\mathbf{X}-\mathbf{D}^{(k)}-\mathbf{E}^{(k)}\right\|_F}{\|\mathbf{X}\|_F}>10^{-8}$ and $\frac{\left\|\hat{\mathbf{D}}^{(k)}-[\mathbf{D}^{(k)};\mathbf{1}^T]\right\|_F}{\left\|\hat{\mathbf{D}}^{(k)}\right\|_F}>10^{-8}$ **do**

Assuming $\mathbf{W}=diag\{w_{ii}\}$, update $\mathbf{T}^{(k+1)}=[\mathbf{t}_1,\mathbf{t}_2,\cdots,\mathbf{t}_c]$, where $\mathbf{t}_i=w_{ii}^2(w_{ii}^2\hat{\mathbf{D}}^{(k+1)}(\hat{\mathbf{D}}^{(k+1)})^T+\gamma\mathbf{I}_d)^{-1}\mathbf{y}_i(\hat{\mathbf{D}}^{(k)})^T$, and γ regularizes the scale of $\mathbf{t}_i$.

Update $\hat{\mathbf{D}}^{(k+1)}=\left[\eta(\mathbf{T}^{(k)})^T\mathbf{W}^T\mathbf{W}\mathbf{T}^{(k)}+\mu_2^{(k)}\mathbf{I}_d\right]^{-1}\left[\eta(\mathbf{T}^{(k)})^T\mathbf{W}^T\mathbf{Y}-\Gamma_2^{(k)}+\mu_2^{(k)}[\mathbf{D}^{(k)};\mathbf{1}^T]\right]$;

Update $\mathbf{D}^{(k+1)}=\mathcal{D}_{1/\beta}(\mathbf{Z}^{(k+1)})$, where $\mathbf{Z}^{(k+1)}=\frac{1}{\beta}\left(\Gamma_1^{(k)}+\mu_1^{(k)}\left(\mathbf{X}-\mathbf{E}^{(k)}\right)+\left[\Gamma_2^{(k)}+\mu_2^{(k)}\hat{\mathbf{D}}^{(k)}\right]_{(1:d_x,\cdot)}\right)$, and $\beta=\mu_1^{(k)}+\mu_2^{(k)}$;

Update $\mathbf{E}^{(k+1)}=\mathcal{S}_{\lambda/\mu_1^{(k)}}\left(\mathbf{X}-\mathbf{D}^{(k)}+\Gamma_1^{(k)}/\mu_1^{(k)}\right)$;

Update $\Gamma_1^{(k+1)}=\Gamma_1^{(k)}+\mu_1^{(k+1)}(\mathbf{X}-\mathbf{D}^{(k+1)}-\mathbf{E}^{(k+1)})$

$\Gamma_2^{(k+1)}=\Gamma_2^{(k)}+\mu_2^{(k+1)}(\hat{\mathbf{D}}^{(k+1)}-[\mathbf{D}^{(k+1)};\mathbf{1}^T])$

$\mu_1^{(k+1)}=\rho\mu_1^{(k)}$

$\mu_2^{(k+1)}=\rho\mu_2^{(k)}$

end while

Output: $\mathbf{T}$, $\mathbf{D}$, $\mathbf{E}$

Algorithm 3 ALM algorithm for cleaning the test data Eq. (3.8)

Input: $\mathbf{X}_{te}\in\mathbb{R}^{d_x\times n_{te}}$, $\mathbf{D}\in\mathbb{R}^{d_x\times n}$, parameters λ (a positive scalar weights term $\|\mathbf{E}\|_1$, which is determined in training) and ρ_t (a positive scalar for updating the Lagrange coefficients).

Initialization:$\mathbf{Z}_{te}^{(0)}=\mathbf{0}_{n\times n_{te}}$, where its element $\mathbf{z}_{te}^{(0)}(i,j)=1$ if $i=arg\min_i\{dist\left(\mathbf{x}_{te}(j),\mathbf{d}_i\right)\}_{i=1,\cdots,n}j=1,\cdots,n_{te}$; $\mathbf{E}_{te}^{(0)}=\mathbf{X}_{te}-\mathbf{D}\mathbf{Z}_{te}^{(0)}$;

Lagrange Multiplier Initialization: $\Gamma_{te}^{(0)}=\frac{\mathbf{X}_{te}}{\|\mathbf{X}_{te}\|_F}$,$\mu_{te}^{(0)}=\frac{dn}{4}\|\mathbf{X}_{te}\|_1$.

while $\frac{\left\|\mathbf{X}_{te}-\mathbf{D}\mathbf{Z}_{te}^{(k)}-\mathbf{E}_{te}^{(k)}\right\|_F}{\|\mathbf{X}_{te}\|_F}>10^{-8}$ **do**

Update $\mathbf{S}^{(k+1)}=\mathbf{Z}_{te}^{(k)}-\frac{1}{\beta_{te}}\left(-\mathbf{D}^T\Gamma^{(k)}+\mu_{te}^{(k)}\mathbf{D}^T\left[\mathbf{D}\mathbf{Z}_{te}^{(k)}-(\mathbf{X}-\mathbf{E}_{te}^{(k)})\right]\right)$, where $\beta_{te}=\mu_{te}^{(k)}\|\mathbf{D}^T\mathbf{D}\|_F^2$;

Update $\mathbf{Z}_{te}^{(k+1)}=\mathcal{D}_{1/\beta}(\mathbf{S}^{(k+1)})$;

Update $\mathbf{E}_{te}^{(k+1)}=\mathcal{S}_{\lambda/\mu_{te}^{(k)}}\left(\mathbf{X}_{te}-\mathbf{D}\mathbf{Z}_{te}^{(k)}+\Gamma_{te}^{(k)}/\mu_{te}^{(k)}\right)$;

Update $\Gamma_{te}^{(k+1)}=\Gamma_{te}^{(k)}+\mu_{te}^{(k)}(\mathbf{X}-\mathbf{D}\mathbf{Z}_{te}^{(k+1)}-\mathbf{E}_{te}^{(k+1)})$, $\mu_{te}^{(k+1)}=\rho_t\mu_{te}^{(k)}$;

end while

Output: $\mathbf{Z}_{te}$, $\mathbf{E}_{te}$

the output labels $\mathbf{Y}$. Our assumption is that the clean test data can be reconstructed as local combinations of the training data, that is $\mathbf{D}_{te} = \mathbf{D}\mathbf{Z}_{te}$, where $\mathbf{Z}_{te} \in \mathbb{R}^{n \times n_{te}}$. In order to make the combination locally compact, we regularize the combination coefficient $\mathbf{Z}_{te}$ by minimizing its nuclear norm [102]. The resulting objective function becomes

$$\begin{aligned} \min_{\mathbf{Z}_{te}, \mathbf{E}_{te}} \quad & \|\mathbf{Z}_{te}\|_* + \frac{\lambda}{\|\mathbf{D}\|_*} \|\mathbf{E}_{te}\|_1 \\ s.t. \quad & \mathbf{X}_{te} = \mathbf{D}\mathbf{Z}_{te} + \mathbf{E}_{te}, \end{aligned} \tag{3.8}$$

where the weight $\frac{\lambda}{\|\mathbf{D}\|_*}$ in front of $\|\mathbf{E}_{te}\|$ is used to keep the original balance between $\|\mathbf{E}_{te}\|$ and $\|\mathbf{D}_{te}\| = \|\mathbf{D}\mathbf{Z}_{te}\|$ in Eq. (3.7). Directly applying the ALM to solve Eq. (3.8) is a challenging task because we cannot apply the standard Singular Value Thresholding (SVT) operators on $\mathbf{Z}_{te}$. Observe that the term $\mathbf{D}\mathbf{Z}_{te}$ is not the standard formulation to be solved with SVT. We followed the idea of [103], and linearized the term $\mathbf{D}\mathbf{Z}_{te}$ before the standard SVT operation. Alg. 3 describes the optimization strategy. After solving (3.8), the regression or classification output for $\mathbf{X}_{te}$ is computed as $\mathbf{Y}_{te} = \mathbf{T}[\mathbf{D}\mathbf{Z}_{te}; \mathbf{1}^T]$. In the case of classification, $\mathbf{Y}_{te}$ is typically used as decision values (for computing AUROC), or to produce binary class labels using the k-nearest-neighbor method.

3.1.5 Robust regression with missing data

Robust regression Eq. (3.3) can be easily extended to handle missing elements in the input data matrix $\mathbf{X}$. From now on, we will refer to this problem as "RR-Missing".

Let Ω be the index set of observed elements in $\mathbf{X}$, and $\mathcal{P}_\Omega$ be the projection operator from the matrix space to the support of observed elements. The RR-Missing solves the following problem

$$\begin{aligned} \min_{\mathbf{T}, \mathbf{D}, \mathbf{E}} \quad & \frac{\eta}{2} \left\| \mathbf{W}(\mathbf{Y} - \mathbf{T}\hat{\mathbf{D}}) \right\|_F^2 + \|\mathbf{D}\|_* + \lambda \|\mathbf{E}\|_1 \\ s.t. \quad & \mathcal{P}_\Omega(\mathbf{X}) = \mathcal{P}_\Omega(\mathbf{D} + \mathbf{E}), \;\; \hat{\mathbf{D}} = [\mathbf{D}; \mathbf{1}^T], \end{aligned} \tag{3.9}$$

The algorithm for solving Eq. (3.9) is similar to Eq. (3.3). After solving Eq. (3.9), the missing elements in $\mathbf{X}$ are filled by the values in $\mathbf{D}$.

As in the case of RR, the test data with missing elements can be cleaned similarly to section 3.1.4 by solving

$$\begin{aligned} \min_{\mathbf{Z}_{te},\mathbf{E}_{te}} \quad & \|\mathbf{Z}_{te}\|_* + \frac{\lambda}{\|\mathbf{D}\|_*}\|\mathbf{E}_{te}\|_1 \\ s.t. \quad & \mathcal{P}_\Omega(\mathbf{X}_{te}) = \mathcal{P}_\Omega(\mathbf{D}\mathbf{Z}_{te} + \mathbf{E}_{te}). \end{aligned} \tag{3.10}$$

After solving Eq. (3.10), the regression/classification output for $\mathbf{X}_{te}$ is computed as $\mathbf{Y}_{te} = \mathbf{T}[\mathbf{D}\mathbf{Z}_{te}; \mathbf{1}^T]$. The extension of RR-Missing to RLDA-Missing is straightforward.

3.1.6 Experimental Results

This section compares our RR methods against state-of-the-art approaches on four experiments for regression and classification.

The first experiment uses synthetic data to compare with existing approaches and illustrate how existing robust regression methods cannot remove outliers that lie in the subspace of the data. The second experiment applies RR to the problem of head pose estimation from partially corrupted images. The third experiment reports comparisons of RR against state-of-the-art multi-label classification algorithms on the MSRC, Mediamill and TRECVID2011 databases. The fourth experiment illustrates the application of RR-Missing to predict facial attributes.

Robust regression (RR) on synthetic data

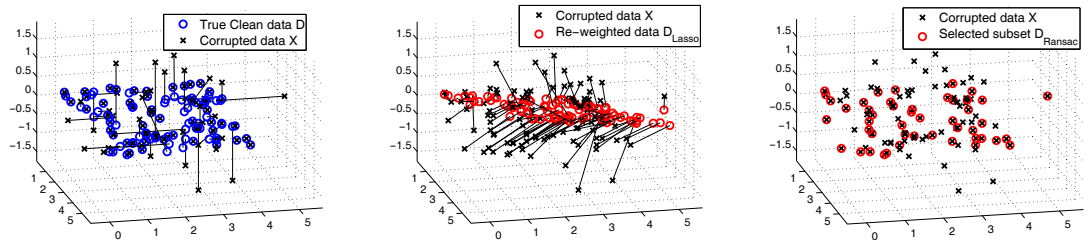
This section illustrates the benefits of RR in a synthetic example. We generated 200 three-dimensional samples, where the first two components were generated from a uniform distribution between $[0, 6]$, and the third dimension is 0. In Matlab notation, $\mathbf{D} = [6 * rand(2, 200); \mathbf{0}^T]$, $\mathbf{X} = \mathbf{D} + \mathbf{E}$, $\mathbf{Y} = \mathbf{T}_*[\mathbf{D}; \mathbf{1}^T]$, where $\mathbf{D} \in \mathbb{R}^{3\times 200}$ is the clean data.

Table 3.1: Relative Absolute Error (RAE) and its standard deviation for output $\mathbf{Y}_{te}$ and regression matrix $\mathbf{T}$ on synthetic data (10 repetitions).

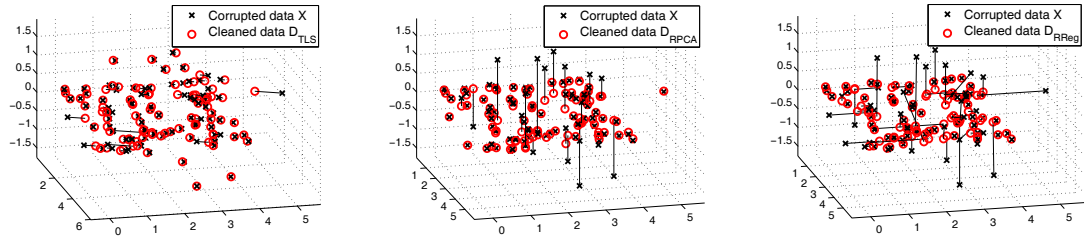
	$RAE_{\mathbf{T}}$	$RAE_{\mathbf{Y}}$
LSR	0.269 ± 0.121	0.035 ± 0.012
GLasso	0.269 ± 0.121	0.035 ± 0.012
RANSAC	0.256 ± 0.133	0.036 ± 0.013
TLS	0.269 ± 0.121	0.925 ± 0.136
RPCA+LSR	0.464 ± 0.030	0.051 ± 0.006
RR	$\mathbf{0.035 \pm 0.015}$	$\mathbf{0.015 \pm 0.006}$

$\mathbf{T}_* \in \mathbb{R}^{3\times 4}$ is randomly generated and used as the true regression matrix. The error term, $\mathbf{E} \in \mathbb{R}^{3\times 200}$, is generated as follows: for 20 random samples, we added random Gaussian noise ($\sim \mathcal{N}(0,1)$) in the second dimension, this simulates in-subspace noise. Similarly, for another 20 random samples, we added random Gaussian noise ($\sim \mathcal{N}(0,1)$) in the third dimension, this simulates noise outside the subspace. The output data matrix is generated as $\mathbf{Y} = \mathbf{T}_*[\mathbf{D}; \mathbf{1}^T] \in \mathbb{R}^{3\times 200}$. Fig. 3.2 (a) shows the clean data $\mathbf{D}$ with blue "o"s, and the corrupted data $\mathbf{X}$ with black "×"s. For better visualization, we only showed 100 randomly selected samples. The black line segments connect the same samples before ($\mathbf{D}$) and after corruption ($\mathbf{X}$). The line segments along the vertical direction are the out-of-subspace component of $\mathbf{E} = \mathbf{X} - \mathbf{D}$, while the horizontal line segments represent the in-subspace component of $\mathbf{E}$.

We compared our RR with five state-of-the-art methods: (1) Standard least-squares regression (LSR), (2) GroupLasso (GLasso) [104], (3) RANSAC [84], (4) Total Least Square (TLS) [105] that assumes the error in the data is additive and follows a Gaussian distribution, (5) RPCA+LSR, which consists of first performing RPCA [50] on the input data, and then learning the regression on the cleaned data using standard LSR. The LSR learns directly the regression matrix $\mathbf{T}$ using the data $\mathbf{X}$. The other methods (2)-(5) re-weight the data or select a subset of the samples input data $\mathbf{X}$ before learning the regression. We randomly select 100 samples for training and the remaining 100 data points for testing. Both the training and testing sets contain half of the corrupted samples. Fig. 3.2(b-f) visualizes the results of the regression for the different methods. Fig. 3.2(b) shows the results of $\mathbf{TX}$,



(a) The clean and corrupted Data (b) Group Lasso (feature re-weighting) (c) RANSAC (subset selection)



(d) TLS (e) RPCA+LSR (f) Robust Regression (RR)

Figure 3.2: (a) Original and corrupted 3D synthetic dataset. Black lines connect data points before (**D**) and after corruption (**X**). (b)-(e) show the input data processed by several baselines, and (f) shows that RR removes the in-subspace outliers.

once **T** is learned with GLasso. GLasso learns a sparse regression matrix that re-weights the input data along dimensions, but it is unable to handle within sample outliers. Observe how the samples are far away from the original clean samples. Fig. 3.2(c) shows the subset of **X** selected by RANSAC. Although we optimized RANSAC's testing sample size to obtain the best testing error measures by RAE, many of the corrupted data points are still identified as inliers. Fig. 3.2 (d) shows results obtained by TLS, where TLS only partially cleaned the corrupted data because the synthesized error cannot be modeled by an isotropic Gaussian distribution. Fig. 3.2 (e) shows results obtained by the method RPCA+LSR, that first computes RPCA to clean the data and then LSR. The data cleaned by RPCA [50], $\mathbf{D}_{RPCA}$, is displayed with red "o"s. Because $\mathbf{D}_{RPCA}$ is computed in an unsupervised manner, only the out-of-subspace error (the vertical lines) can be discarded, while the in-subspace outliers can not be corrected. Finally, Fig. 3.2 (f) shows the result of RR. The clean data $\mathbf{D}_{RR}$ is denoted by red"o"s. Observe that our approach is able to clean both the in-subspace outliers

(the horizontal lines) and out-of-subspace (the vertical lines). This is because our method computes jointly the regression and the subspace estimation.

We also computed the error for the regression matrix $\mathbf{T}_*$ (the first two columns) and the testing error for $\mathbf{Y}_{te}$ on the 100 test samples. Table 3.1 compares the mean regression error measured by the Relative Absolute Error (RAE) between the true labels $\mathbf{Y}_{te} \in \mathbb{R}^{3\times 100}$ and the estimated labels $\widetilde{\mathbf{Y}_{te}}$. $RAE_{\mathbf{T}} = \frac{\|\tilde{\mathbf{T}}(:,1:2)-\mathbf{T}_*(:,1:2)\|_F}{\|\mathbf{T}_*(:,1:2)\|_F}$ and $RAE_{\mathbf{Y}} = \frac{\|\widetilde{\mathbf{Y}}_{te}-\mathbf{Y}_{te}\|_F}{\|\mathbf{Y}_{te}\|_F}$. The information in the third column of $\mathbf{T}_*$ is excluded in generating $\mathbf{Y} = \mathbf{T}[\mathbf{D}; \mathbf{1}^T]$. Therefore, we dismiss this column when evaluating $RAE_{\mathbf{T}}$. As shown in Table 3.1, RR produces the smallest estimation error for both $\mathbf{T}_*$ and $\mathbf{Y}_{te}$ among the five compared methods, while GroupLasso, RANSAC and RPCA+LSR produce small improvements over standard LSR due to their limitation to deal with both the in-subspace and out-of-subspace corruptions.

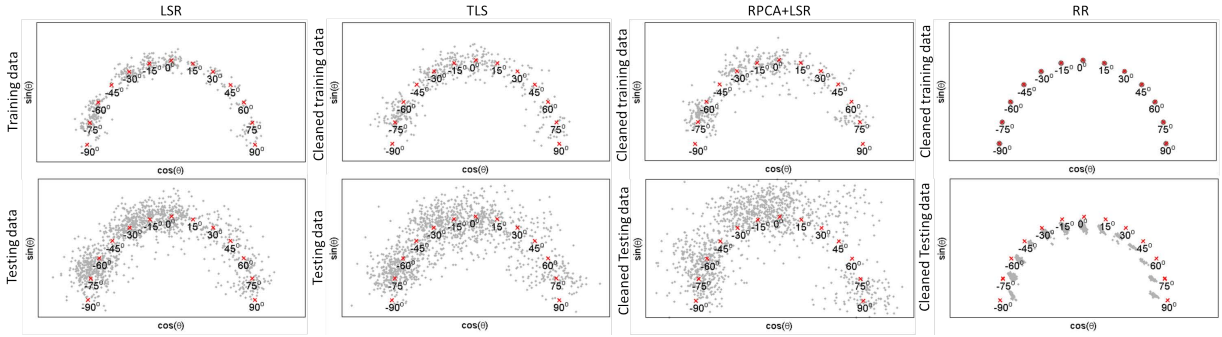


Figure 3.3: Projection of face images (the gray "·"s)in the output space $\mathbf{Y} = [\cos(\theta), \sin(\theta)]$ by LSR, TLS, RPCA+LSR and Robust Regression(RR). The red "×"s denote the ground truth location for pose angles θ =[$-90°$, $-75°$, $-60°$, $-45°$, $-15°$, $0°$, $15°$, $45°$, $60°$, $75°$, $90°$] in the output space.

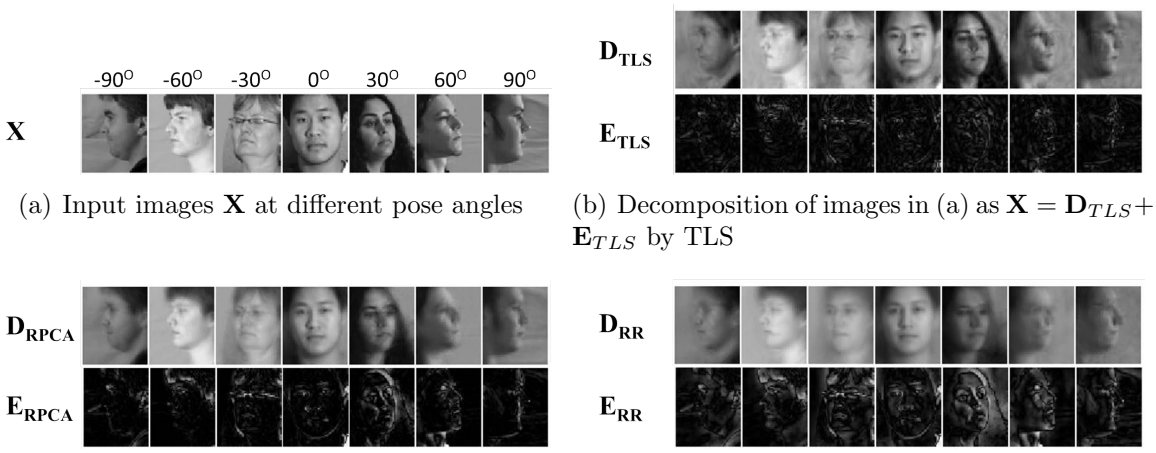
Pose estimation as a RR problem

This section illustrates the benefit of RR in the problem of head pose estimation. We used a subset of CMU Multi-PIE database [106] that contains 1721 face images from 249 subjects in session 1. The face regions are detected automatically using the OpenCV [1] face detector.

[1]http://opencv.willowgarage.com/wiki/

The detected faces cover 11 head poses θ =[$-90°$, $-75°$, $-60°$, $-45°$, $-15°$, $0°$, $15°$, $45°$, $60°$, $75°$, $90°$] each with a random lighting direction. Each image is cropped around the face region and resized to 51×61. We vectorized the images into a vector of $51 \times 61 = 3111$ dimensions in the matrix $\mathbf{X} \in \mathbb{R}^{3111 \times 1721}$ and the yaw angles of the images are used as the output data $\mathbf{Y} = [\cos(\theta), \sin(\theta)] \in \mathbb{R}^{2 \times 1721}$. See Fig. 3.4 for examples of cropped images.

Similar to the previous section, we have compared RR with five methods to learn a regression from the image $\mathbf{X}$ to the yaw angle $\mathbf{Y}$: (1) LSR, (2) GLasso [104], (3) RANSAC [84], (4) TLS and (5) RPCA+LSR. For a fair comparison, we randomly divided the 249 subjects into 5 folds and performed 5-fold cross-validation, at each cross-validation train on 1 fold and test on the remaining 4. Parameters of interest in methods (2)-(4) were selected by performing grid search over the 5-fold cross-validation. The performance of the compared methods is measured with the averaged angle error.



(a) Input images $\mathbf{X}$ at different pose angles

(b) Decomposition of images in (a) as $\mathbf{X} = \mathbf{D}_{TLS} + \mathbf{E}_{TLS}$ by TLS

(c) Decomposition of images in (a) as $\mathbf{X} = \mathbf{D}_{RPCA} + \mathbf{E}_{RPCA}$ by RPCA

(d) Decomposition of images in (a) as $\mathbf{X} = \mathbf{D}_{RR} + \mathbf{E}_{RR}$ by RR.

Figure 3.4: Decomposition of input images in (a) by (b) TLS, (c) RPCA and (d) RR. Robust regression (RR) cleans most facial details and only preserves the correlated with pose angles.

Table 3.2 summarizes the results of methods (1)-(4) and RR. The LSR method produced the largest angle error with the increasing percentage of outliers. RANSAC produced comparable error as standard LSR, indicating that RANSAC is unable to select a subset of

Table 3.2: Comparison of yaw angle error and standard deviation for six methods on a subset of CMU Multi-PIE database [106].

LSR	GLasso	RANSAC	TLS	RPCA+LSR	**RR**
$7.3^o \pm 6.1^o$	$7.1^o \pm 5.9^o$	$7.3^o \pm 6.2^o$	$11.7^o \pm 10.1^o$	$10.8^o \pm 9.7^o$	$\mathbf{5.1}^o \pm \mathbf{4.6}^o$

"inliers" to robustly estimate the regression matrix. RPCA+LSR produced relatively larger yaw angle error. This is because RPCA is unsupervised and lacks the ability to preserve the discriminative information in $\mathbf{X}$ that correlates with the angles $\mathbf{Y}$. RR got the smallest error among all the compared methods.

To further illustrate how RR differs from TLS and RPCA+LSR, Fig. 3.4 visualizes the decomposition of training images by RR (*i.e.*, $\mathbf{X} = \mathbf{D}_{RR} + \mathbf{E}_{RR}$), by TLS (*i.e.*, $\mathbf{X} = \mathbf{D}_{TLS} + \mathbf{E}_{TLS}$) and by RPCA (*i.e.*, $\mathbf{X} = \mathbf{D}_{RPCA} + \mathbf{E}_{RPCA}$), for the same input images. Images under pose angles contains person-specific features *e.g.*, glasses at $-30°$ and long dark hair at $30°$ (see Fig. 3.4(a)). Fig. 3.4(b)-(c) show that both TLS and RPCA are able to remove some of the edges. While RR (Fig. 3.4(d)) preserves much less personal facial details in $\mathbf{D}_{RR}$ than TLS ($\mathbf{D}_{TLS}$) and RPCA ($\mathbf{D}_{RPCA}$) (especially images under pose $-30°$ and $30°$). With less facial details and more dominant profiles, the regression trained on $\mathbf{D}_{RR}$ (as in RR) is able to model higher correlation with the pose angles than using $\mathbf{D}_{RPCA}$.

Fig. 3.3 visualizes the differences among LSR, TLS, RPCA+LDA and RR on both training (the 1^{st} row) and testing images (the 2^{nd} row). We projected the face images (the gray "$\cdot$"s) into the output space $\mathbf{Y} = [\cos(\theta), \sin(\theta)]$ using the discussed four methods (one column each). The red "$\times$"s denote the ground true location for pose angles. The projections (the gray "$\cdot$"s) produced by LSR, TLS and RPCA+LDA are far from the ideal outputs (the red "$\times$"s). RR (the 4^{th} column) is the method that improves the correlation between inputs (the gray "$\cdot$"s) and the outputs (the red "$\times$"s), therefore is more robust than LSR, TLS and RPCA+LSR in estimating the pose angles.

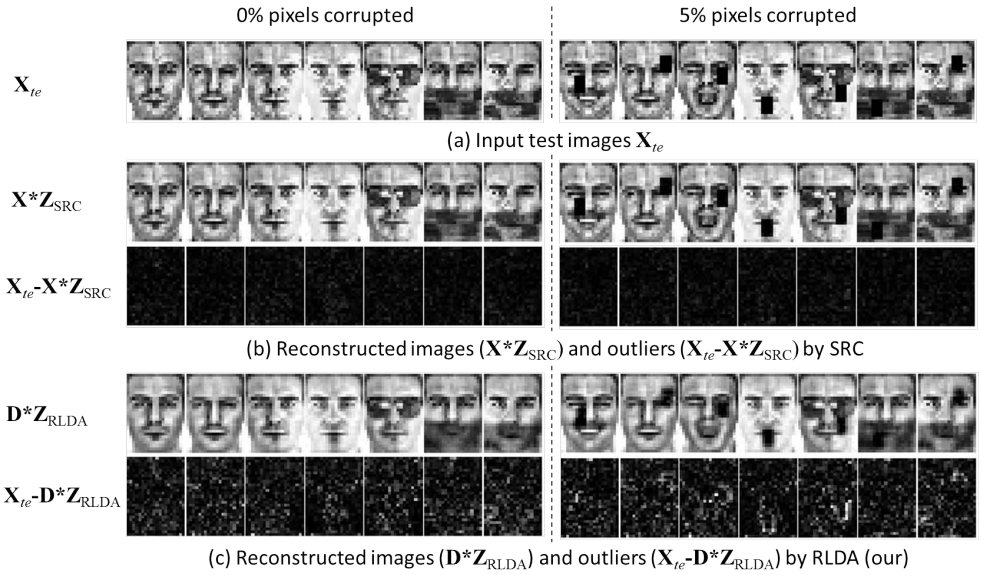


Figure 3.5: Decomposition of downsampled test images $\mathbf{X}_{te}$ in the AR face database [107]. **Left**: Experiments on original images (0% corruption). **Right**: Experiments on synthetically corrupted images (5% corruption). (a) Input test images; (b) Reconstructed test images ($\mathbf{X}\mathbf{Z}_{SRC}$) and the outliers ($\mathbf{X}_{te} - \mathbf{X}\mathbf{Z}_{SRC}$) by Sparse Representation for Classification (SRC) [108], where $\mathbf{X}$ is the training images and $\mathbf{Z}_{SRC}$ is the sparse coefficient for the test images $\mathbf{X}_{te}$; (c) Reconstructed test images ($\mathbf{D}\mathbf{Z}_{RLDA}$) and the outliers ($\mathbf{X}_{te} - \mathbf{D}\mathbf{Z}_{RLDA}$) by Robust LDA (RLDA), where $\mathbf{D}$ is the cleaned training images by solving Eq. 3.6, and $\mathbf{Z}_{RLDA}$ is the RLDA coefficient computed by Eq. 3.8. Observe that RLDA cleaned more intra-sample outliers and reconstruct more facial details than SRC.

Robust LDA (RLDA) for face recognition

This section evaluates our Robust LDA (RLDA) method for face recognition with synthetically corrupted images.

We used the AR database [107]. There are over $4,000$ frontal face images of 126 subjects under illumination change, expressions, and facial disguises. 26 pictures were taken for each subject and organized in two sessions. In the experiment, we used the cropped and aligned face images provided in [107]: 50 male subjects and 50 female subjects. For each subject, 13 images from Session 1 were used for training, and the rest 13 images from Session 2 for testing. Each image was cropped and resized to 165×120 and converted to gray-scale (see

the first row on the left of Fig. 3.5 for examples). To evaluate robustness of algorithms, we randomly corrupted the image by replacing image pixels using black squares (see the first row on the right half of Fig. 3.5 for examples).

We followed the settings in [108], and used two types of features that produced the highest performance of Sparse Representation for Classification (SRC) in [108]: (1) Downsampled face: downsample the cropped images by 1/6, and vectorize a downsampled image into a 540 dimensional vector; (2) Laplacian face: compute Laplacian face features [109] on the original 165×120 image and select the top 540 components. Fig. 3.5 illustrates decomposition of downsampled test images $\mathbf{X}_{te}$ (a) in the AR face database [107] by SRC and our Robust LDA (RLDA) approach. The **Left** part of Fig. 3.5: Experiments on original images (0% corruption). The **Right** part of Fig. 3.5: Experiments on synthetically corrupted images(5% corruption). Using SRC [108] (Fig. 3.5(b)), the test images were reconstructed as $(\mathbf{XZ}_{SRC})$, where $\mathbf{X}$ is the training images and $\mathbf{Z}_{SRC}$ is the sparse coefficient. The outliers was then computed as $(\mathbf{X}_{te} - \mathbf{XZ}_{SRC})$. Observe that SRC produced little outliers. This is because both the training and testing images of the same subject contain similar expression, illumination, glasses and scarf. SRC computed the sparse representation of test images $\mathbf{X}_{te}$ using similar training images in $\mathbf{X}$. Fig. 3.5(c) shows the reconstructed test images $(\mathbf{DZ}_{RLDA})$ and the outliers $(\mathbf{X}_{te} - \mathbf{DZ}_{RLDA})$ by RLDA, where $\mathbf{D}$ is the cleaned training images by solving Eq. 3.6, and $\mathbf{Z}_{RLDA}$ is the RLDA coefficient obtained by Eq. 3.8. Note different to SRC, our RLDA approach used the cleaned training images $\mathbf{D}$ instead of the original training images $\mathbf{X}$. We can see from Fig. 3.5(c) that RLDA cleaned more intra-sample outliers and reconstruct more facial details than SRC.

In Table 3.3, we compared face recognition accuracy of linear SVM, SRC and RLDA using the both the downsampled images and the Laplacian face as classification features. As shown in the first row (0%) in Table 3.3 , RLDA produced the higher accuracy than SRC and SVM on downsampled images, and comparable accuracy to SRC on Laplacian

features. From the 2^{nd} to 4^{th} row, with the higher corruption, all methods showed lower accuracy. Furthermore, because the Laplacian features were not computed in the robust manner, under high corruptions (the 3^{rd} to 4^{th} row in Table 3.3), the results with Laplancian features were worse than RLDA with the downsampled images. Comparing to SVM and SRC, RLDA showed the best robustness for it consistently produced the best results.

Table 3.3: Face recognition accuracy on AR face database [107] under synthetic corruption. The percentages in the brackets denotes the portion of images covered by the synthetic squares. *Higher* value indicates better performance. Best results are in bold.

%-pixel corruption	1-NN	SVM	SRC	RLDA
Downsample (0%)	68.5%	76.4%	88.0%	**89.8**%
Laplacian (0%)	90.8%	80.6%	**94.7**%	**94.8**%
Downsample (5%)	33.7%	64.7%	80.8%	**85.1**%
Laplacian (5%)	54.5%	74.5%	71.7%	**77.2**%
Downsample (20%)	9.7%	44.5%	67.5%	**72.4**%
Laplacian (20%)	47.8%	**67.9**%	63.6%	64.9%
Downsample (40%)	7.4%	35.5%	52.9%	**61.4**%
Laplacian (40%)	33.7%	**56.5**%	48.2%	51.4%

RLDA for object classification, action recognition and video indexing

This section evaluates our Robust LDA (RLDA) method on two multi-label and one multi-class classification tasks: object categorization on the MSRC dataset, action recognition in the MediaMill dataset and event video indexing on the TRECVID 2011 dataset. Each dataset corpus and features is described below:

MSRC Dataset (Multi-label)[2] has 591 photographs (see Fig. 3.6(a)) distributed among 21 classes, with an average of 3 classes per image. We mimic [74] and divide each image into an 8×8 grid and calculate the first and second order moments for each color channel on each grid in the RGB space. This results in a 384 dimensional vector, which we use to describe each image.

Mediamill Dataset (Multi-label) [110] consists of 43907 sub-shots (see Fig. 3.6(c)) divided in 101 classes. We followed [74] and eliminated classes containing fewer than 1000 samples, leaving 27 classes. Then, we randomly selected 2609 sub-shots such that each

[2]http://research.microsoft.com/en-us/projects/ObjectClassRecognition/

class has at least 100 labeled data points. Each image was therefore characterized by a 120-dimensional feature vector, as described in [110].

PASCAL VOC 2007 Dataset (Multi-label) consists of 9963 images labeled with at least one of 20 classes, split into `trainval` and `test` sets. We used state of the art features obtained from Overfeat, a Convolutional Neural Network trained on ImageNet [111]. We rescaled every image to 221×221 pixels and obtained a single 4096 dimensional feature vector as the output from layer 22 of the network for every image in the dataset.

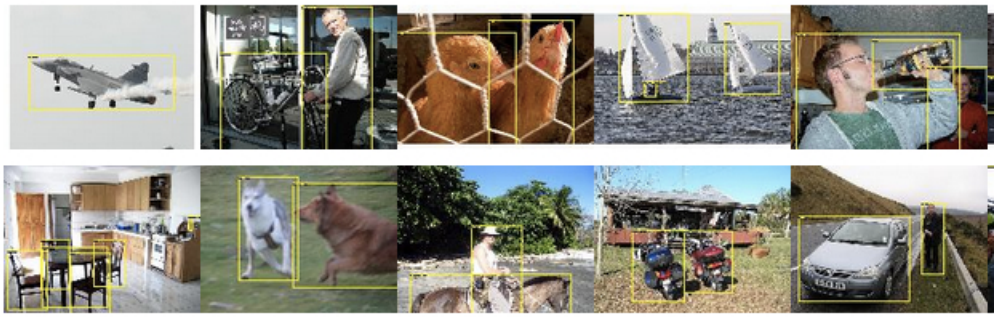
TRECVID 2011 Dataset (Multi-class)[3] consists of video data in MED 2010 and the development data of MED 2011, totaling 9822 video clips belonging exclusively to one of 18 classes. We first detected 100 shots for each video and then used their center frames as keyframes. We described each keyframe using dense SIFT descriptors. From these, we learned a 4096 dimension Bag-of-Words dictionary. Each video was represented by a normalized histogram of all of its feature points. We used a 300 core cluster to extract the SIFT features, which took about 1500 CPU hours in total. In the experiment, we randomly split the dataset into two subsets: 3122 entries for training and 6678 for testing.

We compared RLDA to the state of the art approach for Multi-Label LDA (MLDA) [74], and to Robust PCA [50] followed by traditional LDA (RPCA+LDA). For control, we also compare to LDA, PCA+LDA (preserving 99.9% of energy) and a linear one-*vs.*-all SVM.

For the classic LDA-based testing procedure, one first projects the test points using the learned $\mathbf{T}$ from training; then for each projected test sample, we find the k-nearest-neighbor (kNN) from the training samples projected by $\mathbf{T}$; finally, we select the class label from the class labels of k-neighbors by majority voting. However, this procedure is not appropriate in our evaluation for two reasons: (1) it's not fair to use a fixed k for classes with different number of samples, *e.g.*, samples per class are in $[19, 200]$ for MSRC, $[100, 2013]$ for Mediamill; (2) kNN introduces nonlinearity to the LDA-based classifiers, which is unfair to linear SVM.

[3] http://www-nlpir.nist.gov/projects/tv2011/

(a)



(b)

(c)

Figure 3.6: Multi-label datasets for object recognition and action classification. Example images in (a) MSRC and (b) PASCAL VOC 2007 , and (c) example keyframes in Mediamill.

For these reasons, we use Area Under Receiver Operating Characteristic (AUROC) as our evaluation metric. AUROC summarizes the cost/benifit ratio over all possible classification thresholds. We report the average AUROC (over 5-fold Cross Validation) for each method under their best parameters in Table 3.4. In the MSRC dataset results in Table 3.4, LDA performs the worst since it's most sensitive to the noise in data. SVM performs better than PCA+LDA and RPCA+LDA. Our method (RLDA) leads to significant improvements over the others due to its joint classification and data cleaning (for both Gaussian and sparse noise in the input). For Mediamill, LDA is just slightly worse than PCA+LDA and RPCA+LDA due to the low noise level in the data. In this case, RLDA does not "over-clean" the data, and performs similar to PCA+LDA and RPCA+LDA. In the PASCAL VOC 2007 dataset results, performance increases become less accentuated, with baseline methods yielding good performance due to the recent advances in representation provided by Overfeat [111]. MLDA, on the other hand, results in a poorer score because it heavily relies on the normalization based on inter-class correlations.

Table 3.4: AUROC for Multi-label Object (MSRC), Action (Mediamill) and Image (Pascal VOC) classification. *Higher* value indicates better performance. Best results are in bold.

Database	LDA	SVM	PCA+LDA	MLDA	RPCA+LDA	RLDA
MSRC	0.65	0.79	0.76	0.63	0.75	**0.83**
Mediamill	**0.77**	0.64	**0.77**	0.67	**0.77**	0.76
Pascal VOC2007	0.92	0.90	0.92	0.79	0.87	**0.94**

To test our method in a large scale dataset, we run experiments on the TRECVID2011 dataset. We used the Minimum Normalized Detection Cost (MinNDC), the evaluation criteria for MED 2010 and MED 2011 challenges, as suggested by NIST. Fig. 3.7 shows that RLDA achieved the best class-wise MinNDC for 9 out of 18 classes over other linear methods, *i.e.*, LDA/MLDA, SVM and RPCA+LDA. Note because the classes are mutually exclusive, MLDA is identical to LDA. SVM is heavily affected by outliers for the "Wedding Ceremony", "Getting a vehicle unstuck" and "Making a sandwich" cases. For some classes, LDA and RPCA+LDA are similar or better than RLDA. We believe this is due to: (1) the data features

computed by Bag-of-Words model smoothed/regularized some outliers; (2) the nonlinear nature of the classification task. Therefore some error patterns modeled by LDA and RPCA enhanced their discriminative ability. Nevertheless, among all linear algorithms, our method (RLDA) obtains the best average MinNDC. In addition, to show how nonlinearity affects the performances, we compared the kernelized version of the LDA (KLDA), RPCA+LDA (KRPCA+KLDA) and RLDA (KRDA). Here we apply the homogeneous kernel maps technique [112] to obtain a three order approximation of the χ^2 kernel. Other more accurate approximations are possible [113]. Fig. 3.7 shows that KRDA still obtains better results, 9 out of 18 best class-wise MinNDC and best average MinNDC over all classes.

Event Description \ Methods	LDA/MLDA	SVM	RPCA+LDA	**RLDA**	KLDA	KRPCA+KLDA	**KRLDA**
Maling a cake	1.0027	1.0038	0.999	**0.9091**	0.9819	0.9716	**0.9706**
Batting a run	**0.6987**	1.0019	0.9498	0.8552	**0.7413**	0.9931	**0.7416**
Assembling a shelter	0.9989	1.0152	1.0026	**0.9787**	1.0038	**0.938**	0.9994
Attempting a board trick	**1.0019**	**1.0018**	1.0057	**1.0019**	0.9494	**0.882**	0.9876
Feeding an animal	1.0038	**0.9899**	1.0038	1.0019	**0.9889**	1.0076	0.9988
Landing a fish	0.9605	1.0019	0.9169	**0.9056**	0.8937	**0.9941**	**0.8796**
Wedding ceremony	0.9967	12.4498	**0.9789**	0.9923	**0.8048**	0.9787	0.8075
Woodworking project	1.0051	**0.8588**	1.0038	1.0057	1.0032	**0.8156**	1.0019
Birthday party	0.9862	0.9561	**0.9368**	0.9881	**0.9654**	0.9848	0.9703
Changing a vehicle tire	0.9856	0.9842	1.0019	**0.9549**	0.923	0.9855	**0.9223**
Flash mob gathering	0.8384	0.9675	0.8933	**0.8189**	0.7905	0.9936	**0.7167**
Getting a vehicle unstuck	0.9848	11.6719	**0.9659**	0.9867	0.9524	1.0076	**0.9146**
Grooming an animal	**0.9691**	1.0019	0.9868	1.0094	0.9918	**0.8956**	0.9924
Making a sandwich	1.0019	4.0583	1.0132	**0.981**	0.9936	**0.9658**	0.9792
Parade	0.9931	**0.9723**	0.9931	0.9805	1.0006	**0.9877**	1.0038
Parkour	0.9837	1.0019	**0.9336**	1.0019	0.8412	1.0056	**0.8385**
Repairing an appliance	0.9369	**0.5998**	1.0075	0.9344	0.9312	**0.8698**	0.9344
Working on a sewing project	1.0057	1.0056	1.0025	**0.9192**	0.9349	1.0019	**0.9148**
Average Score	0.9641	2.3635	0.9775	**0.9571**	0.9273	0.9601	**0.9208**

Figure 3.7: MinNDC results for Media Event Detection on TREC2011. *Lower* value indicates better performance. Best results are in bold.

RLDA with missing data

This section illustrates the use of RLDA-Missing to perform attribute classification on the PubFig database [114]. We predict 7 facial attributes (Gender, Asian, White, Indian, Black, Glasses and Beard/Mustache) from facial features. That is, we formulate the facial attribute detection as a multi-label classification problem, each image has 7 attribute labels, where a binary indicator vector $\mathbf{y}_i \in \mathbb{R}^{7\times 1}$, such that $y_{ij} = 1$ if $\mathbf{x}_i$ belongs to class j otherwise $y_{ij} = 0$

($j = 1, \cdots, 7$). In testing (see section 3.1.4), a testing data point $\mathbf{x}_{te}$ is cleaned to produce $\mathbf{d}_{te}$ and the indicator vector $\mathbf{y}_{te} = \mathbf{T}[\mathbf{d}_{te}; 1] \in \mathbb{R}^7$, which then is used as decision values (for computing AUROC), or produces binary class labels using the k-nearest-neighbor method.

To train our facial attribute detector, we used training images from the PubFig database that have been labeled with 49 landmarks and images from Multi-PIE database [106] that have been labeled with 68 landmark points. This is a challenging problem because the regressor will have input features of different dimensions. In this section we will show how RR is able to merge information from these two databases to get improved results on estimating facial attributes.

The PubFig database [114] consists of $58,797$ images of 200 people collected from the internet. Classifiers will be trained to recognize these facial attributes from image features. The images in the PubFig database are taken in completely uncontrolled situations with non-cooperative subjects. Thus, there are large variations in pose, lighting, expression, occlusion, scene and camera parameters. These imaging conditions pose great difficulties in classifying the facial attributes. Besides the PubFig database, we also used 5683 face images from the Multi-PIE database. These images include all 249 subjects under the frontal lighting and yaw angle between $-45°$ and $+45°$.

Given the images that have been labeled with the seven attributes, we compute the image features as follows. First, we used the supervised descent method [115] to detect 49 facial landmarks in the PubFig database. Second, we compute a 8-dimensional Histogram of Gradient (HoG) vector around each facial point, (the size of each pixel block is 1/6 of the length of the nose). Finally, we concatenate all the point HoGs to form a $8 \times 49 = 392$-dimensional feature vector for the image. See Fig. 3.8 for an example. In the case of the Multi-PIE images, faces have been manually labeled and we proceed as before and extract a 544 feature vector, see Fig. 3.8 (b).

As a baseline experiment, we applied the RLDA proposed in section 3.1.3, using only

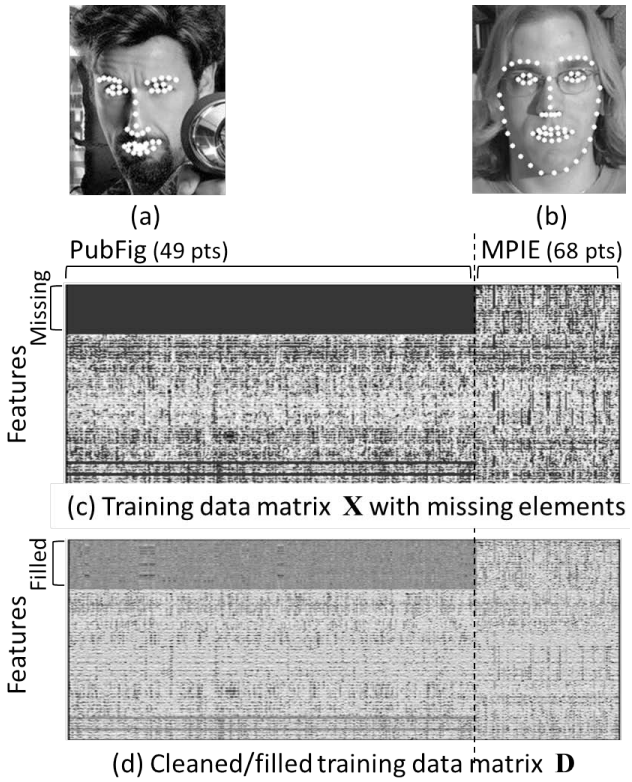


Figure 3.8: Training RLDA-Missing classifier on a concatenated data matrix $\mathbf{X}$ consisting data from the (a) PubFig database (49 facial points detected) and (b) the MultiPIE database (68 facial points detected). In the original concatenated matrix "$\mathbf{X}$" (c), observe that the data block of PubFig contains missing elements. In the clean/filled data matrix "$\mathbf{D}$" (d), the missing elements are automatically filled. In testing, we only use the PubFig part of $\mathbf{D}$ to clean the testing data.

data from the PubFig database. In this experiment, we perform grid-search for RR parameters (η and λ) with a 4-fold cross validation. At each trial of cross-validation we left one fold (50 persons) out for testing, and the rest three folds (images of 150 persons) for training. The averaged AUROCs over 4-fold cross-validation are reported at the optimal parameters (Table. 3.5 the "PubFig only" row). We added data from MultiPIE database to conduct a"PubFig&MultiPIE" experiment. At each trial of cross-validation we added the 5683 MultiPIE images (544-dimensional features) to the 3 PubFig training folds (392-dimensional features). The $544 - 392 = 52$-dimensional unavailable features in the PubFig dataset are considered as missing data (see Fig. 3.8 (c) for the concatenated training data

Table 3.5: AUROCs of facial attribute classification on the PubFig data. Each row contains results using different method and training data, as specified in the first column "Methods: training data". *Higher* value indicates better performance. Best results are in bold.

Methods: training data \ Attributes	Gender	Asian	White	Indian	Black	Glasses	Beard	Average
RLDA: PubFig (49pts) only	**0.92**	0.50	0.60	0.50	**0.72**	0.77	0.68	0.67
RLDA:PubFig (49pts)&MPIE (49pts)	0.90	0.62	0.57	**0.60**	0.69	0.76	0.70	0.69
LDA-missing[116]: PubFig (49pts)&MPIE(68pts)	0.82	0.57	0.54	0.59	0.61	0.78	0.67	0.65
RLDA-missing: PubFig (49pts)&MPIE(68pts)	0.91	**0.66**	**0.70**	0.56	0.69	**0.81**	**0.71**	**0.72**

matrix "**X**"). We train RLDA with missing data as in section 3.1.5, the missing elements in "**X**" is filled in the cleaned/filled training data the "**D**" (Fig. 3.8 (d)). Finally in testing, we only used the PubFig part of "**D**" to clean the testing data (the remaining 1 PubFig fold). All quantitative results are shown in Table. 3.5. In addition to the "PubFig only" baseline, we added one more baseline "PubFig (49pts)&MPIE (49pts)" by using features from the 49pts that are common to both datasets. We also implemented the discriminatively trained LDA for missing data in [116], a standard LDA-based approach for missing data. The approach used the same missing training data as RLDA-missing, and the results were reported in Table. 3.5 ("LDA-missing"). Comparing to the two baseline methods ("RLDA: PubFig only" and "RLDA: PubFig (49pts)&MPIE (49pts)"), our RLDA-missing approach have additional 52-dimensional features learned from the MultiPIE data. Comparing to "LDA-missing" [116], our approach does not rely on explicit assumption on the missing values. "LDA-missing" [116] explicitly models the missing values by Gaussian distribution, whereas the missing elements in this experiment are structured (blocked). As shown in Table. 3.5, our RLDA-missing produced improved results in both class-wise and average AUROCs.

3.2 Weakly supervised learning as a matrix completion problem

Most methods for visual recognition are fully supervised in nature, as they make use of bounding boxes or pixelwise segmentations to locate objects of interest. A major limitation of these approaches, however, is that the location for objects of interest has to be known in the training images, usually in the form of bounding boxes or a full-blown pixelwise segmen-

(a) Training (b) Testing

Figure 3.9: In a weakly-supervised method for multi-label image classification, the training set images (a) are labeled with the objects that are present but their location in the image is unknown. Given unseen test images (b), our method is able to classify which classes are present in the image and segment the image into regions that correspond to the classes.

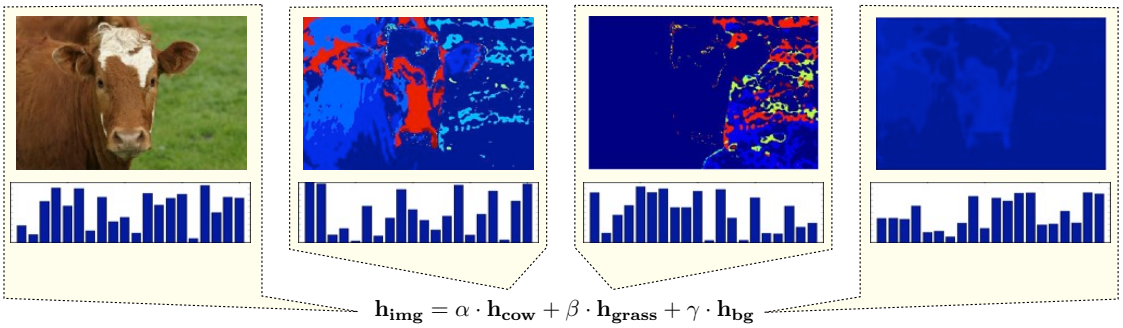


Figure 3.10: The left image represents the original training image that has been labeled with the words grass and cow. Our algorithm decomposes the histogram of this image as a linear combination of two class histogram basis (cow, grass) plus another histogram $\mathbf{h_{bg}}$ modeling errors and the background. Class localization can be visualized on the image by interpreting each histogram as a probability distribution of which words belong to the class.

tation. While efforts have been made to provide datasets with this information [117, 118], manual labeling is still labor intensive, subjective and an error prone process. Moreover, it has been shown that manual segmentations are not necessarily the optimal spatial enclosure for object classifiers [119]. To cope with an increasing number of concepts and larger scale datasets, there has been an increased interest in transitioning away from these fully supervised approaches.

Weakly-supervised algorithms [119, 120, 121, 122] relieve the labeling burden by learning from labels with less information. Figure 3.9 illustrates this setting and the problem we address in this section: given a weakly-labeled training set (Figure 3.9(a)), we segment and label new test images (Figure 3.9(b)). To solve this problem, we propose to cast it under

a matrix completion framework. Several Multiple Instance Learning (MIL) methods [119, 123, 124, 125, 126, 127, 128] have been proposed in the literature for solving this type of weakly supervised learning problem. However, existing MIL methods have four major drawbacks: (1) The MIL problem is usually cast as a NP-hard binary quadratic problem. Consequently, most existing algorithms to solve MIL are highly sensible to initialization. (2) The extension of MIL to the multi-label case is not trivial. Current multi-label MIL approaches [124, 125, 126] heavily rely on an explicit enumeration of instances, which are then solved by single class MIL or Multi-label learning. (3) They lack robustness to outliers. Recall from Section 3.1 that most discriminative approaches project data directly onto linear or non-linear spaces [119, 123, 127]. Thus, a single outlier can bias the solution, severely degrading classification performance. (4) It is unclear how existing MIL approaches can be extended to use partial information, such as incomplete label assignments or missing feature descriptions.

We observe that the image classification and localization problem has more structure than what's exploited by MIL problems. MIL approaches consider images as bags with many instances denoting possible regions of interest. Instead, we make use of the additive property of histogram representations such as Bag of Words (BOW) [129]: the histogram of an entire image is a sum of the histogram information of all of its subparts (see Figure 3.10). By using this property, image classification can be posed as a low-rank matrix completion problem, since class histograms are shared across images, and the number of class histograms is small compared to the number of images. Contrary to typical MIL approaches, our matrix completion model is convex. Figure 3.11 illustrates the main idea in this section. Each column of $\mathbf{Z}^{obs}$ has a concatenation of the labels (1 if the class is present and zero otherwise) and the histogram $\mathbf{h}_i^{tr}$ for one training image (Figure 3.11 (a)). In the test set (Figure 3.11 (b)), labels are unknown and denoted as question marks (?). Our method fills the unknown entries and corrects known features and labels such that $\mathbf{Z}$ has the smallest rank possible. It

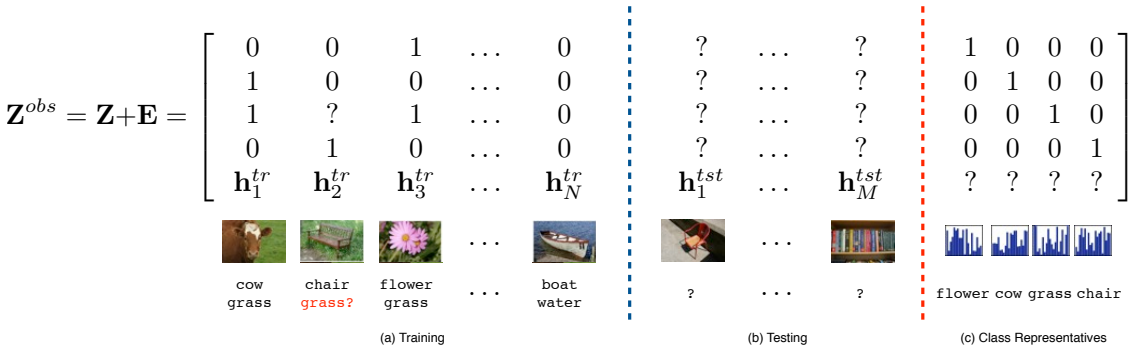


Figure 3.11: Our weakly supervised classification algorithm works by completing a matrix $\mathbf{Z}^{obs}$ as shown above, where the question marks denote unknown entries. We complete this matrix such that it can be factorized into a low rank matrix $\mathbf{Z}$ and an error matrix $\mathbf{E}$. This ensures that background distributions and feature/label outliers are captured in $\mathbf{E}$, since they increase the rank of $\mathbf{Z}$. In the training submatrix (a), the i^{th} column concatenates the image histogram $\mathbf{h_i^{tr}}$ with its respective $\{0, 1\}$ label assignments. Note that a partially labeled example such as the second training image (a) is trivially handled by our framework. In the test submatrix (b), the j^{th} column is a concatenation of histogram $\mathbf{h_j^{tst}}$ with unknown assignments. In this transductive setting, the statistics of the test set are also used in the learning. By completing (c), we obtain a representative histogram for each class, in spite of their co-occurrence in the images.

can also infer the feature descriptor of a particular class (Figure 3.11 (c)). This is achieved by looking for the unknown histograms whose label vector denotes the presence of only this class. In doing so, we obviate the need for training with precise localization or expensive combinatorial MIL models, as required by previous methods.

3.2.1 Related work

Since the seminal work of Barnard and Forsyth [130], many researchers have addressed the problem of associating words to images. Image semantic understanding is now typically formulated as a multi-label problem. In this setting, each image may be simultaneously assigned to more than one class. This is an important difference from multi-class classification, where classes are assumed to be independent and mutually exclusive. While multi-label can trivially be handled in multi-class approaches by dropping the mutual exclusivity constraint, Desai *et al.* [131] have shown the need to model object interactions. Therefore, many multi-

class techniques such as SVM and LDA have been modified to make use of label correlations to improve multi-label classification performance [132]. In these approaches, localization is achieved by detection, using *e.g.*, a sliding window. This is, however, at the expense of a fully supervised training set where localization is known a priori.

Several researchers have addressed the problem of classifying an image and providing precise class localization. Deselaers *et al.* [133] used a CRF to learn new class appearances from previously known ones obtained with supervised training. Blaschko *et al.* [134] learned a supervised structured output regression where the outputs are coordinates of a bounding box enclosing the object. Jamieson *et al.* [122] associated configurations of SIFT features to captions. Tighe and Lazebnik [135] proposed lazy learning for large scale image parsing.

Alternatively to these approaches, Multiple Instance Learning (MIL) has surfaced as a reliable framework for performing learning in the presence of unknown latent factors. First proposed in [136], this class of learning problems extends the typical classification setting to the case where labels are no longer applied individually, but to multi-sets or "bags": a bag is labeled positive if at least one of its instances is positive and negative if none of its constituents are. In computer vision, this framework has been used for weakly supervised learning tasks such as learning deformable part models [127] and to explicitly model the relations between labels and specific regions of the image, as initially proposed by Maron and Lozano-Perez [137].

This method allows for the localization and classification tasks to benefit from each other, thus reducing noise in the corresponding feature space and making the learned semantic models more accurate [39, 119, 123, 124, 125, 126, 138]. Although promising, the MIL framework is combinatorial, so several approaches have been proposed to avoid local minima and deal with the prohibitive number of possible subregions in an image. Zha *et al.* [125] made use of hidden CRFs while Vijayanarasimhan *et al.* [126] recurred to multi-set kernels to emphasize instances differently. Yang *et al.* [123] exploited asymmetric loss functions to balance

false positives and negatives. These methods, however, require an explicit enumeration of instances in the image. This is usually obtained by breaking images in a small fixed number of segments or applied in settings where detectors perform well, such as the problem of associating faces to captioned names [139]. On the other hand, to avoid explicitly enumerating the instances, Nguyen *et al.* [119] coupled constraint generation algorithms with a branch and bound method for fast localization. This is also seen in the negative data-mining process of [127]. Yakhnenko *et al.* [39] proposed a MIL algorithm of linear complexity in the number of instances by using a non-convex Noisy-Or model. Multi-task learning has also been proposed as a way to regularize the MIL problem to avoid local minima due to many available degrees of freedom. In this setting, the MIL optimization is jointly learned with a fully supervised task [138].

To the best of the authors' knowledge, the only work modeling MIL as a convex problem is by Li and Sminchisescu [128]. They replace the classifier loss and the non-convex constraints on the positive bags by convex alternatives (f-divergence family loss and class likelihood ratios for each instance). They show promising results over standard MIL formulations as the ratio of positive instances in positive bags increase. Unfortunately, this is not the setting in image classification, as the percentage of possible negative bounding boxes in an image largely exceeds that of the positives. This work can also relate to Latent Semantic Analysis, as the low rank justifications provided in Sec. 3.2.2 are similar in nature to the ones provided for subspaces obtained from document-term matrices. Bosch *et al.* [140] provided preliminary results that visual words associated with high probability to a given category can provide cues for localization.

3.2.2 Matrix completion for multi-label classification of visual data

This subsection describes the main contributions of this section: We start by presenting the use of matrix completion for general classification tasks. Then, we describe its use for weakly

supervised multi-label image classification and localization.

Classification as a matrix completion problem

In a supervised setting, a classifier learns a mapping $\mathcal{W} : \mathcal{X} \rightarrow \mathcal{Y}$ between the space of features $\mathcal{X}$ and the space of labels $\mathcal{Y}$. This learning is done from a dataset of N training tuples $(\mathbf{x}_i^{tr}, \mathbf{y}_i^{tr}) \in \mathbb{R}^D \times \mathbb{R}^K$, where D is the feature dimension and K the number of classes. In particular, linear classifiers minimize a loss $l(\cdot)$ between the output space and the projection of the input space, as

$$\underset{\mathbf{W},\mathbf{b}}{\text{minimize}} \sum_{i=1}^{N} l\left(\mathbf{y}_i^{tr}, [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{x}_i^{tr} \\ 1 \end{bmatrix}\right), \tag{3.11}$$

where parameters $\mathbf{W} \in \mathbb{R}^{D \times K}$ and $\mathbf{b} \in \mathbb{R}^{K \times 1}$ describe the class decision boundaries. After the training stage, these parameters are used to estimate unknown labels for test samples $\mathbf{y}_j^{tst}$ from their feature descriptors $\mathbf{x}_j^{tst}$. This is typically done independently for each test entry, as

$$\mathbf{y}_j^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{x}_j^{tst} \\ 1 \end{bmatrix}. \tag{3.12}$$

In this section, we formulate the problem of jointly classifying M test samples as one of matrix completion. For this purpose, let us define the feature matrices $\mathbf{X}^{tr} \in \mathbb{R}^{D \times N}$ and $\mathbf{X}^{tst} \in \mathbb{R}^{D \times M}$. These matrices respectively collect in each column feature vectors for N training and M test samples. Without loss of generality, the linear model assumed in (3.12) can be written in matrix form. Specifically, it states that $\mathbf{Y}^{tr} \in \mathbb{R}^{K \times N}$, the matrix concatenating the labels for all training images, can be obtained by the linear combination

$$\mathbf{Y}^{tr} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tr} - \mathbf{E}_X^{tr} \\ \mathbf{1}^\top \end{bmatrix} - \mathbf{E}_Y^{tr}, \tag{3.13}$$

where $\mathbf{E_Y}^{tr}$ and $\mathbf{E_X}^{tr}$ denote errors in the known labels and features, respectively. The test labels $\mathbf{Y}^{tst} \in \mathbb{R}^{K \times M}$ are obtained as

$$\mathbf{Y}^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tst} - \mathbf{E}_X^{tst} \\ \mathbf{1}^\top \end{bmatrix}, \tag{3.14}$$

with no error in the labels since they are unknown. Concatenating labels and features in (3.13) and (3.14) in one matrix yields

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X^{tr} & \mathbf{E}_X^{tst} \\ \mathbf{0}^\top \end{bmatrix} = \mathbf{Z}^{obs} - \mathbf{E}, \tag{3.15}$$

where $\mathbf{Z}^{obs} \in \mathbb{R}^{(K+D+1)\times(M+N)}$ holds all observed entries (with $\mathbf{Y}^{tst}$ unknown) and $\mathbf{E}$ is a matrix of errors, also unknown. Note that according to (3.13) and (3.14), the matrix $\mathbf{Z}$ defined in (3.15) is rank deficient. That is, the rows comprising the labels are linearly dependent on the feature rows. In the absence of error ($\mathbf{E} = \mathbf{0}$), the input matrix $\mathbf{Z}^{obs}$ is also low rank, as

$$\operatorname{rank}(\mathbf{Z}) = \operatorname{rank}(\mathbf{Z}^{obs}) = \operatorname{rank}\left(\begin{bmatrix} \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix}\right). \tag{3.16}$$

In this case, we observe that (3.13) becomes

$$\mathbf{Y}^{tr} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tr} \\ \mathbf{1}^\top \end{bmatrix}, \tag{3.17}$$

and thus the $\mathbf{Y}^{tst}$ in (3.14) does not increase the rank of $\mathbf{Z}$, since

$$\mathbf{Y}^{tst} = [\mathbf{W}^\top \ \mathbf{b}] \begin{bmatrix} \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix}. \tag{3.18}$$

Using this result, Goldberg *et al.* [49] suggested that unknown test labels in $\mathbf{Y}^{tst}$ can be recovered by completing these entries such that the rank of $\mathbf{Z}$ is minimized. This can be written as

$$\begin{aligned} \underset{\mathbf{Y}^{tst}}{\text{minimize}} \quad & \text{rank}(\mathbf{Z}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top & \end{bmatrix}. \end{aligned} \tag{3.19}$$

In practice $\mathbf{E} \neq \mathbf{0}$, so we modify (3.19) to include (3.15). To avoid trivial solutions, we penalize errors with a loss $l(\cdot)$, as

$$\begin{aligned} \underset{\mathbf{Y}^{tst}, \mathbf{E}_Y^{tr}, \mathbf{E}_X}{\text{minimize}} \quad & \text{rank}(\mathbf{Z}) + \lambda l(\mathbf{E}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top & \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X & \\ \mathbf{0}^\top & \end{bmatrix}, \end{aligned} \tag{3.20}$$

where λ is a tradeoff parameter and $\mathbf{E}_X = [\mathbf{E}_X^{tr} \ \mathbf{E}_X^{tst}]$. We discuss the choices of loss functions $l(\cdot)$ in detail in Sec. 3.2.2.

Since the rank is a highly non-convex and non-differentiable function, it is nontrivial to minimize. Therefore, we relax (3.20) by using its convex envelope, the nuclear norm.

Therefore, we rewrite (3.20) as

$$\begin{aligned} \underset{\mathbf{Y}^{tst},\mathbf{E}_{\mathbf{Y}^{tr}},\mathbf{E}_X}{\text{minimize}} \quad & \|\mathbf{Z}\|_* + \lambda l(\mathbf{E}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top & \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X & \\ \mathbf{0}^\top & \end{bmatrix}. \end{aligned} \tag{3.21}$$

There are three fundamental advantages in casting a general classification problem as the matrix completion in (3.21). First, it bypasses the estimation of the model parameters $\mathbf{W}$ and $\mathbf{b}$. This allows our formulation to estimate errors in the features $\mathbf{E}_X$. Parametric models that estimate $\mathbf{W}$ and $\mathbf{b}$ (such as linear regression or SVMs) do not model this error, and thus implicitly assume $\mathbf{E}_X = \mathbf{0}$. Note that the product $\mathbf{W}^\top \mathbf{E}_X$ in (3.13) will result in a non-convex problem when both $\mathbf{W}$ and $\mathbf{E}_X$ are considered as optimization variables, whereas our model (3.21) is convex. Second, errors and missing data in features and labels are estimated jointly. Third, we minimize the rank of $\mathbf{Z}$, containing training and test samples. This transductive setting allows the model to leverage the statistics of the test set.

Adding robustness into matrix completion

In practical applications, we have several sources of errors in the features (*e.g.*, changes in pose, illumination, background noise) and missing data in the training samples (*e.g.*, missing labels), which will translate into nonzero error matrices in the models of (3.15) and (3.35). We account for these possible violations by allowing the matrix $\mathbf{Z}$ in (3.21) to deviate from the original data matrix. The resulting optimization problem finds the best label assignment $\mathbf{Y}^{tst}$ and error matrices $\mathbf{E}_X = [\mathbf{E}_{X^{tr}} \ \mathbf{E}_{X^{tst}}], \mathbf{E}_{Y^{tr}}$ such that the rank of $\mathbf{Z}$ is minimized, as

$$\begin{aligned} \underset{\mathbf{Y}^{tst},\mathbf{E}_{Ytr},\mathbf{E}_X}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + l_x(\mathbf{E}_X) + \lambda l_y(\mathbf{E}_Y^{tr}) \\ \text{subject to} \quad & \mathbf{Z} = \begin{bmatrix} \mathbf{Y}^{tr} & \mathbf{Y}^{tst} \\ \mathbf{X}^{tr} & \mathbf{X}^{tst} \\ \mathbf{1}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{E}_Y^{tr} & \mathbf{0} \\ \mathbf{E}_X \\ \mathbf{0}^\top \end{bmatrix}. \end{aligned} \tag{3.22}$$

Here, distortions of $\mathbf{Z}$ from known labels and features are penalized according to $l_y(\cdot)$ and $l_x(\cdot)$, respectively. The parameters λ, μ are positive trade-off weights between better feature adaptation and label error correction. We rewrite (3.22) by defining sets Ω_X and Ω_Y of known feature and label entries and $\mathbf{Z}_Y, \mathbf{Z}_X, \mathbf{Z}_1$ as the label, feature and last rows of $\mathbf{Z}$, as

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\sum_{ij\in\Omega_X} l_x(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|}\sum_{ij\in\Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \qquad , \\ \text{subject to} \quad & \mathbf{Z}_1 = \mathbf{1}^\top \end{aligned} \tag{3.23}$$

where the constraint that $\mathbf{Z}_1$ be equal to one is necessary for dealing with the bias $\mathbf{b}$ in (3.13). The model in (3.23) can be solved using Fixed Point Continuation [55], described in Sec. 3.2.3.

In [49], $l_x(\cdot)$ was defined as the least squares error and $l_y(\cdot)$ a log loss to emphasize the error on entries switching classes as opposed to their absolute numerical difference. We note that in this model (MC-1), the log loss in $l_y(\cdot)$, albeit asymmetric, incurs in unnecessary penalization of entries belonging to the same class as the original entry (see Figure 3.12). Therefore, we generalize this loss to a smooth approximation of the Hinge loss, controlled by a parameter γ. For labels $\{-1, 1\}$, we have

$$l_y(z_{ij}, z_{ij}^{obs}) = \frac{1}{\gamma}\log(1 + \exp(-\gamma z_{ij}^{obs} z_{ij})), \tag{3.24}$$

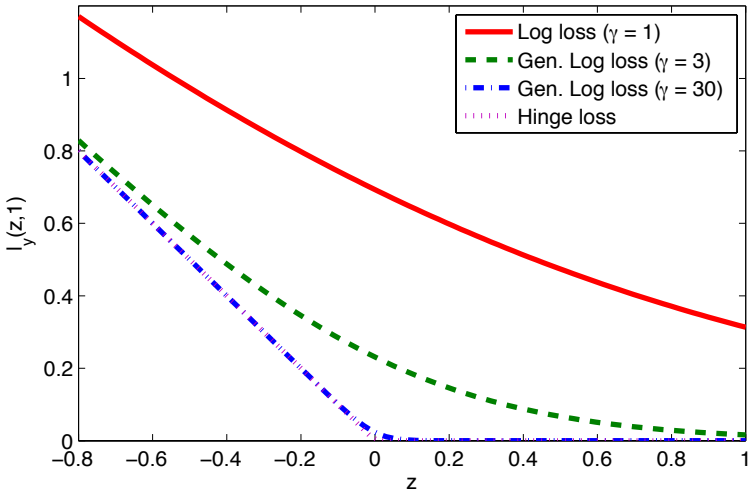


Figure 3.12: Comparison of Generalized Log loss with Log loss ($\gamma = 1$).

and for the case of labels $\{0, 1\}$, we have

$$l_y(z_{ij}, z_{ij}^{obs}) = \frac{1}{\gamma} \log\left(1 + \exp\left(-\gamma(2z_{ij}^{obs} - 1)(z_{ij} - z_{ij}^{obs})\right)\right). \tag{3.25}$$

Also, in the bag of words model, visual data are encoded as histograms. In this setting, (3.23) is inadequate as it introduces negative values to the histograms in $\mathbf{Z}_X$. Thus, we replace the least-squares penalty in $l_x(\cdot)$ by a χ^2 distance,

$$\chi^2(\mathbf{z}^j, \mathbf{z}_0^j) = \sum_{i=1}^{F} \chi_i^2(z_{ij}, z_{ij}^{obs}) = \sum_{i=1}^{F} \frac{\left(z_{ij} - z_{ij}^{obs}\right)^2}{z_{ij} + z_{ij}^{obs}}. \tag{3.26}$$

and constrain all feature vectors to be positive

Definition 6 (MC-Pos model)**.**

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|} \sum_{ij \in \Omega_X} \chi_i^2(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|} \sum_{ij \in \Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \\ \text{subject to} \quad & \mathbf{Z}_X \geq \mathbf{0} \\ & \mathbf{Z}_1 = \mathbf{1}^\top 61 \end{aligned} \tag{3.27}$$

or in the Probability Simplex $\mathcal{P}$

Definition 7 (MC-Simplex model)**.**

$$\begin{aligned} \underset{\mathbf{Z}}{\text{minimize}} \quad & \mu\|\mathbf{Z}\|_* + \frac{1}{|\Omega_X|}\sum_{ij\in\Omega_X}\chi_i^2(z_{ij}, z_{ij}^{obs}) \\ & + \frac{\lambda}{|\Omega_Y|}\sum_{ij\in\Omega_Y} l_y(z_{ij}, z_{ij}^{obs}) \\ \text{subject to} \quad & \mathbf{Z}_X \in \mathcal{P} \\ & \mathbf{Z}_1 = \mathbf{1}^\top, \end{aligned} \tag{3.28}$$

depending on whether we wish to perform normalization on the data or not. Observe that for the presented losses $l_x(\cdot)$ and $l_y(\cdot)$, (3.23) and (3.28) are both convex in their domains.

3.2.3 Fixed point continuation (FPC) for MC-Pos/MC-Simplex

Albeit convex, the nuclear norm makes (3.27) and (3.28) not smooth. Since nuclear norm problems are naturally cast as Semidefinite Programs, existing interior point methods are inapplicable due to the large dimension of $\mathbf{Z}$. Thus, several methods have been devised to efficiently optimize this problem class [22, 53, 54, 55, 56, 57, 58]. The FPC method [55], in particular, consists of a series of gradient descent updates $h(\cdot) = I(\cdot) - \tau g(\cdot)$ with step size τ and gradient $g(\cdot)$ as

$$g(z_{ij}) = \begin{cases} \frac{\lambda}{|\Omega_Y|}\frac{-z_{ij}^{obs}}{1+\exp(\gamma z_{ij}^{obs} z_{ij})} & \text{if } z_{ij} \in \Omega_Y, \\ \frac{1}{|\Omega_X|}\frac{z_{ij}^2 + 2z_{ij}z_{ij}^{obs} - 3{z_{ij}^{obs}}^2}{(z_{ij}+z_{ij}^{obs})^2} & \text{if } z_{ij} \in \Omega_X, \end{cases} \tag{3.29}$$

and 0 otherwise. These steps are alternated with a shrinkage operator $S_\nu(\cdot) = \max(0, \cdot - \nu)$ on the singular values of the resulting matrix, to minimize its rank. Provided $h(\cdot)$ is non-expansive, FPC converges to the optimal solution for the unconstrained problem. FPC was originally devised in [55] for unconstrained problems and extended in [49] to solve the formulation MC-1 (3.23) by adding a projection step. However, its convergence to the global optima of the problem was only empirically verified. In Appendix B, we prove the convergence of FPC for (3.23), (3.27), (3.28) using the fact that projections onto convex sets are non-expansive.

Key to the feasibility of FPC is an efficient way to project $\mathbf{Z}$ onto the constraint sets in (3.27) and (3.28). While for MC-Pos (3.27) the non-negative orthant projection is done by setting negative components to zero, efficiently projecting onto the probability simplex in MC-Simplex (3.28) is not straightforward. By exploring the dual of the projection problem we obtain a closed form, cf. [62, 141]. The algorithms are summarized in Alg. 4 and we prove their convergence in Appendix B. The computational bottleneck is the computation of the SVD of $\mathbf{Z}$. State-of-the-art methods for SVD (*e.g.*, Lanczos bidiagonalization with partial reorthogonalization) take a flop count of $O((K + D + 1)(M + N)^2 + (M + N)^3)$.

Algorithm 4 FPC for MC-Pos (3.27) and MC-Simplex (3.28)

Input: Initial Matrix $\mathbf{Z}^{obs}$, known entries sets Ω_X, Ω_Y
  Initialize $\mathbf{Z}$ as the rank-1 approximation of $\mathbf{Z}^{obs}$
  for $\mu = \mu_1 > \mu_2 > \cdots > \mu_k$ **do**
    while Rel. Error $> \epsilon$ **do**
      Gradient Descent: $\mathbf{A} = \mathbf{Z} - \tau g(\mathbf{Z})$
      Shrink 1: $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$
      Shrink 2: $\mathbf{Z} = \mathbf{U}S_{\tau\mu}(\mathbf{\Sigma})\mathbf{V}^\top$
      Project $\mathbf{Z}_X$: $\mathbf{Z}_X = \max(\mathbf{Z}_X, \mathbf{0})$ for (3.27)
      Project $\mathbf{Z}_X$ onto the probability simplex $\mathcal{P}$ for (3.28)
      Project $\mathbf{Z}_1$: $\mathbf{Z}_1 = \mathbf{1}^\top$
    end while
  end for
Output: Complete Matrix $\mathbf{Z}$

3.2.4 Low-rank assumption for histograms in image classification

In spite of justifying the applicability of matrix completion as a generic classification framework, the explanation provided by Goldberg *et al.* [49] described in Sec. 3.2.2 only spans the row space of $\mathbf{Z}$. In this section, we provide an alternative explanation for the low rank of $\mathbf{Z}$ in (3.15), based instead on its column space. Let us assume the case when histograms are used as feature vectors. Note that several popular techniques for obtaining global representations of images in computer vision, such as Bag of Words or HOG, fall under this assumption. Let $\mathbf{h}^i$ denote such a histogram representation for image i. In this case, the feature submatrix $\mathbf{X} = [\mathbf{X}^{tr}\ \mathbf{X}^{tst}]$ in (3.15) contains one histogram per column, as

$$\mathbf{X} = \left[\begin{array}{ccc|ccc} \mathbf{h}_1^{tr} & \cdots & \mathbf{h}_N^{tr} & \mathbf{h}_1^{tst} & \cdots & \mathbf{h}_M^{tst} \end{array}\right]. \tag{3.30}$$

One property of image histograms is that they can be represented by a sum of the histograms of its segments. We consider these latent histograms as $\mathbf{C}_k \in \mathbb{R}^{D \times N_k}$, the N_k canonical histogram representations for class k. Therefore, we have that the histogram of image i can be written as a sum of class representatives $\mathbf{C}_k$ weighted by coefficients $\mathbf{a}_{k,i} \in \mathbb{R}^{N_k \times 1}$, as

$$\mathbf{h}_i = \sum_k \mathbf{C}_k \mathbf{a}_{k,i} + \mathbf{E}_{Xi}, \tag{3.31}$$

where $\mathbf{E}_{Xi}$ collects errors (*e.g.*, words in the background that do not pertain to any class). If we concatenate the representatives $\mathbf{C}_k$ in the matrix

$$\mathbf{C} = \left[\begin{array}{ccccc} \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_K \end{array}\right], \tag{3.32}$$

and collect weights $\mathbf{a}_{k,i}$ in a matrix $\mathbf{A}$ we can write (3.30) as

$$\mathbf{X} = \mathbf{C}\mathbf{A} + \mathbf{E}_X. \tag{3.33}$$

Additionally, since we postulated each $\mathbf{c}_{kj}$ as belonging to only class k, the correspondent label matrix for $\mathbf{C}$ is given by

$$\mathbf{Y}_C = \begin{bmatrix} \mathbf{e}_1 \mathbf{1}_{N_1}^\top & \cdots & \mathbf{e}_K \mathbf{1}_{N_K}^\top \end{bmatrix}, \tag{3.34}$$

where $\mathbf{e}_i$ denotes the i^{th} canonical vector. Merging (3.30) and (3.34), we obtain the data matrix $\mathbf{Z}^{obs}$ in (3.15) as

$$\mathbf{Z}^{obs} = \begin{bmatrix} \mathbf{Y}_C \\ \mathbf{C} \end{bmatrix} \mathbf{A} + \begin{bmatrix} \mathbf{E}_Y \\ \mathbf{E}_X \end{bmatrix} = \mathbf{Z} + \mathbf{E}, \tag{3.35}$$

the sum of a low rank component matrix $\mathbf{Z}$ with an error matrix $\mathbf{E}$. A close inspection of (3.35) allows us to state that $\mathbf{Z}^{obs}$ is low rank also due to its column space, in absence of background noise, since class histograms are shared across images and therefore $\sum_k N_k < N + M$. Additionally, it allows for the observation that the appearance of individual classes can be recovered from a multi-label dataset by estimating $\mathbf{C}$. In this chapter, we assume that for localization purposes, each class can be well represented by a single histogram. In this case, (3.34) becomes $\mathbf{Y}_C = \mathbf{I}_K$, and therefore our approach can obtain an estimate of $\mathbf{C}$ by completing in $\mathbf{Z}^{obs}$ the features correspondent to the canonical labels (see Figure 3.11 (c)). By directly estimating $\mathbf{C}$, we are able to recover the appearance of each class and thus provide the localization for each concept in the images. This is done despite the weakly supervised setting and bypassing the combinatorial nature of searching for bounding boxes such as in MIL problems. Also, note that this assumption is not used in the classification, where our algorithm estimates class subspace dimensions automatically.

Low rank assumption validation

We empirically validated the low-rank assumption that histograms of objects of the same class share a low-dimensional subspace in two multi-label datasets, MSRC[4] and SIFTFlow [142]. We constructed a bag of words representation for the MSRC dataset, which consists of 591 real world images distributed among 21 classes, with an average of 3 classes present per image. To replicate the setup of [125, 126], we dismissed the classes void, mountain and horse. To obtain a bag of words (BOW) descriptor, we clustered texton filter responses [143] obtained from all three CIELab color channels into a codebook by applying k-means to a random subset of $40,000$ descriptors. In this model [129], images are encoded as histograms representing the distribution of the 400 words from the codebook. Then, using the ground truth segmentation labeling, we collected feature matrices $\mathbf{X}_1$ by concatenating all the histograms of the same class. We compared these with feature matrices $\mathbf{X}_2$ of the same dimension with an equal amount of elements from all classes (including elements from the class of $\mathbf{X}_1$). In order to compare the singular value distribution of these matrices, we normalized them so columns have unit ℓ_2 norm. Then, we measured their nuclear norm ratio (NNR), defined as

$$NNR(\mathbf{X}_1, \mathbf{X}_2) = \frac{\|\mathbf{X}_1\|_*}{\|\mathbf{X}_2\|_*}. \tag{3.36}$$

This measure provides an empirical validation of our assumption and is linked to what our model is optimizing and is an indirect measure of the rank of a matrix. Results on Table 3.6(a) show that for all classes in the MSRC dataset, we obtained a NNR lower than 1. An assignment of test entries to incorrect class labels yields a higher nuclear norm of $\mathbf{Z}$, thus validating our model. For visualization, we plot the singular value distribution of $\mathbf{X}_1$ and corresponding $\mathbf{X}_2$ for some classes in the dataset (Figure 3.13).

It might be argued that explanation of (3.35) only holds when the columns dominate the

[4]http://research.microsoft.com/en-us/projects/objectclassrecognition/

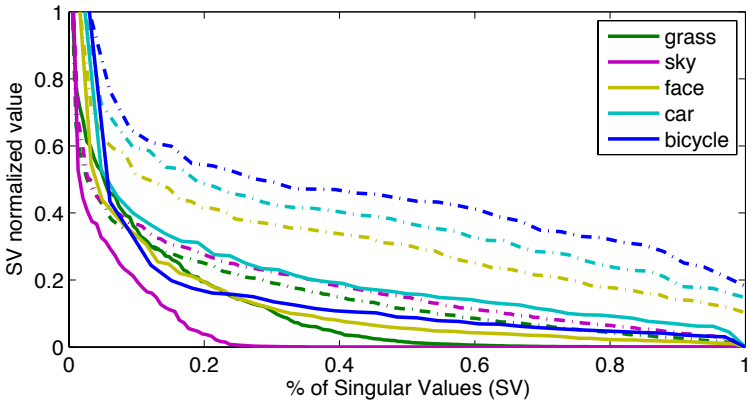


Figure 3.13: Comparison of singular value distribution of matrices $\mathbf{X}_1$ with histograms of the same class (solid) versus corresponding matrices $\mathbf{X}_2$ of the same dimension with an equal amount of histograms from all classes (dashed) for different classes on the MSRC dataset.

estimate of the rank, *i.e.*, $\text{rank}(\begin{bmatrix} \mathbf{X}^{tr} & \mathbf{X}^{tst} \end{bmatrix}) \leq N + M \leq F$. However, we also validated this hypothesis in the case when the feature dimension F is smaller than the number of images $N + M$ in the dataset. Since there are only 591 images in the MSRC dataset and some classes exhibit a small number of exemplars, we validated this assumption in the larger scale SIFTFlow dataset [142]. This dataset is a collection of $2,688$ images distributed among 33 classes. Following [142], we extracted a dense HoG feature map [144] from every image in the dataset and built a BoW codebook of 200 words. We collected the histograms for all the $25,758$ ground truth segments in the dataset according to their class label. Then, we calculated the distribution of singular values for matrices $\mathbf{X}_1$ as aforementioned, for all classes with more than 200 samples in the dataset. We compared the NNR of the matrices $\mathbf{X}_1$ with matrices $\mathbf{X}_2$ of the same dimension comprised by an equal amount of elements from all classes. Results in Table 3.6(b) corroborate the MSRC dataset results, showing our assumption is also valid when the feature dimensions are smaller than the number of images.

3.2.5 Comparison to other subspace techniques

It is important to note that many standard dimensionality reduction techniques such as PCA and LDA have been robustified by using a nuclear norm penalization typically coupled with

Table 3.6: Nuclear norm ratios (NNR) for all classes in the MSRC dataset (a) and for all classes which have more than 200 segments in the SIFTflow dataset (b).

(a) MSRC dataset.

Class	NNR
building	0.8595
grass	0.6987
tree	0.8325
cow	0.9092
sheep	0.7653
sky	0.4530
aeroplane	0.7831
water	0.8224
face	0.6622
car	0.8392
bicycle	0.6525
flower	0.8741
sign	0.9491
bird	0.8793
book	0.9217
chair	0.9397
road	0.7070
cat	0.8402
dog	0.8420
body	0.9465
boat	0.9123

(b) SIFTflow dataset.

Class	NNR
building	0.9074
tree	0.8620
car	0.8989
sky	0.8455
window	0.7513
mountain	0.8657
road	0.8568
person	0.8673
plant	0.8655
sidewalk	0.9038
rock	0.8728
door	0.6554
sea	0.6073
field	0.7719
sign	0.9098
grass	0.8181
streetlight	0.9439
river	0.8719
balcony	0.7458

an ℓ_1 error function [44, 145]. The differences and similarities between the method presented in Section 3.2.2 and these techniques can be analyzed if one interprets (3.27),(3.28) as forms of PCA with missing data. Our method can be seen as an extension of Robust PCA in two ways: 1) it includes labels as additional "features" in the data samples 2) it penalizes label and features errors with different losses l_x and l_y.

A comparison between the behavior of PCA, LDA, RPCA [145], RLDA [44] and our method in the presence of noise can be seen in Figure 3.14. We generated a two-class dataset of $2,000$ 500-dimensional vectors. The positive and negative classes (resp.) have $1,000$ samples of the form $-\mathbf{1}_{500}$ and $\mathbf{1}_{500}$ (resp.). We refer to this as clean data. The first two principal components of this clean data are in Figure 3.14(a). Then, we added to the clean data noise sampled from a Normal distribution with zero mean and standard deviation $20\mathbf{I}_{500\times 500}$. We plot the two principal components data in Figure 3.14(b). Note that PCA does not recover the underlying structure of the clean data due to the significant amount of

noise.

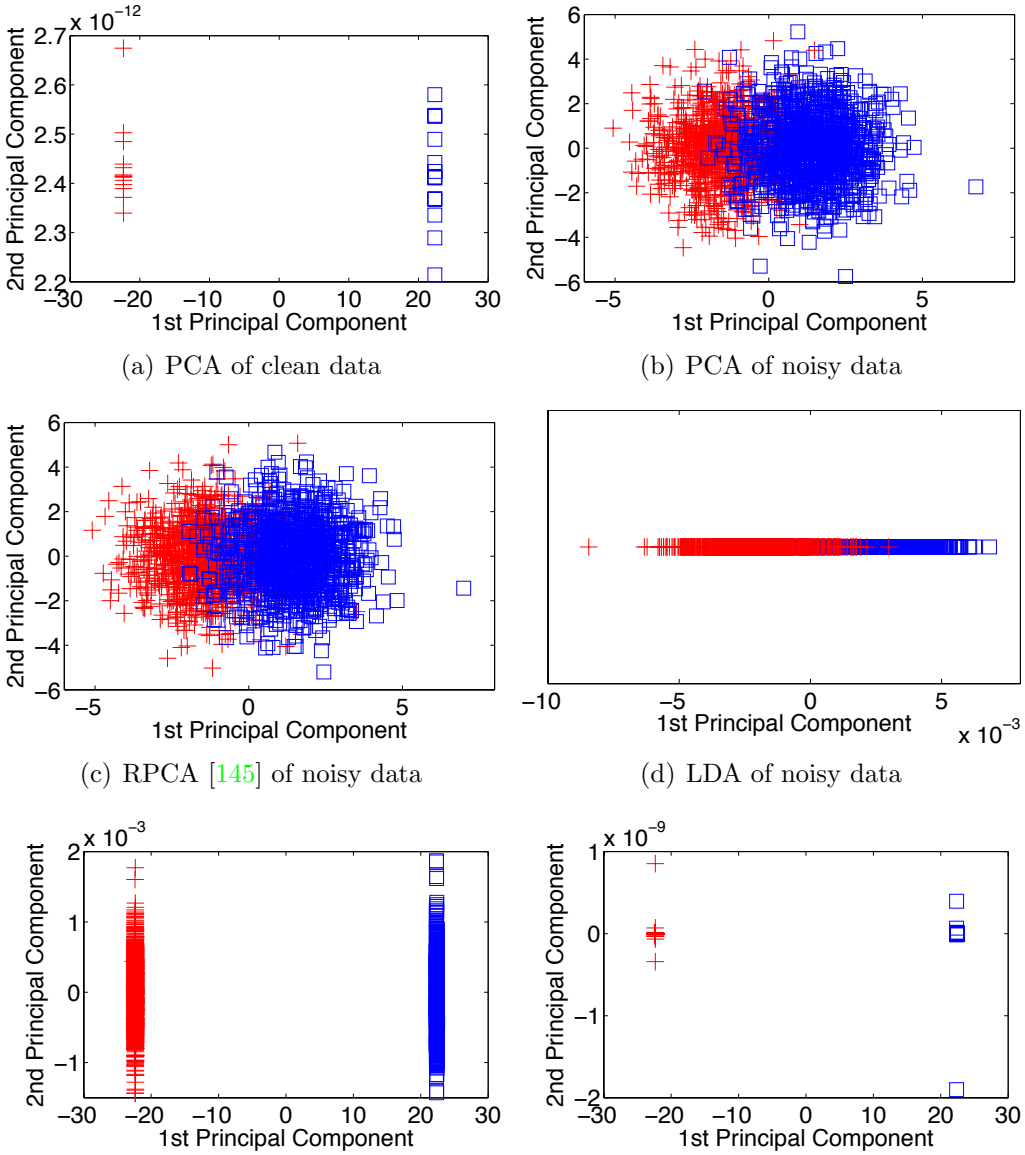
In this example, because the data does not have outliers and the noise does not follow a Laplacian distribution, the ℓ_1 error function assumed by RPCA [145] is not able to clean the noisy data (Figure 3.14(c)). Similarly, augmenting the space by adding the labels as an additional dimension does not help since for RPCA the errors in features and labels are weighted equally. In both these cases, the output of RPCA (Figure 3.14(c)) is similar to the one obtained by regular PCA (Figure 3.14(b)). LDA (Figure 3.14(d)) is able to find a projection which classifies most of the points correctly. However, observe that it fails to clean the data, which results in several misclassified points on the class boundary. Our matrix completion approach, in turn, balances a trade-off between correcting the data points, correcting the labels and minimizing the rank. Therefore, it is able to correct the feature data (Figure 3.14(e)) by giving more weight to the information on the labels. This capability of correcting the errors in features is only matched by our work in Robust LDA [44], which achieved the result in Figure 3.14(f). While this method has the advantage of obtaining an explicit transformation from the feature to the label space, the matrix completion has the ability to clean the test data during training.

3.2.6 Experimental results

This section presents the evaluation of the algorithms MC-Pos (3.27) and MC-Simplex (3.28) in several tasks. In Sec. 3.2.6, we evaluated the classification and localization performance of our method on the CMU-Face dataset [119] (a two-class problem). Second, we evaluated the performance of our method for multi-label classification in the MSRC and PASCAL VOC2007 datasets. Lastly, we also perform an experiment for localization in MSRC.

Parameters

For MC-Pos, MC-Simplex and MC-1, the values considered for parameter tuning were $\gamma \in [1, 30], \lambda \in [10^{-4}, 10^{2}]$. The continuation steps require a decreasing sequence of μ, which we



(a) PCA of clean data

(b) PCA of noisy data

(c) RPCA [145] of noisy data

(d) LDA of noisy data

(e) PCA of MC-1 ($\lambda = 100$)

(f) RLDA [44] of noisy data

Figure 3.14: Comparison of results obtained for two-class classification of the random dataset in 3.2.5. Unlike others, the error correction in Robust LDA (f) and Matrix Completion (e) allow for recovery of the original data.

chose as $\mu_k = 0.25\mu_{k-1}$, stopping when $\mu = 10^{-9}$. We used $\mu_0 = 0.25\sigma_1$, where σ_1 is the largest singular value of $\mathbf{Z}^{obs}$, with unknown entries set to zero. Convergence was defined as a relative change in the objective function smaller than 10^{-2}. In a transduction setting, since the task is to classify an already known test set, one could choose the parameters which perform best on the final test set. However, to be fair to other baselines, we tuned the parameters in a cross validation setting. As such, the results reported are for the choice of parameters which, from the aforementioned ranges, yielded the best average result on all the validation sets provided by cross-validation. The results reported for the SVM baselines were obtained using libSVM, with parameter $C \in [10^{-6}, 10^{6}]$.

Classification and localization on a two-class problem

In this experiment, we tested the classification performance of our method in a two-class classification problem. We used the CMU Face dataset [146], which consists of 624 images of 20 subjects. All subjects are captured with varying expression and poses, with and without sunglasses. Figure 3.15 shows examples of our positive (wearing sunglasses) and negative class (not wearing sunglasses). We have two goals: First, we want to build a classifier that, given a new face image, determines whether the subject is wearing sunglasses or not. Second, Nguyen *et al.* [119] argue that better results are obtained when the classifier training is restricted to the region that has the discriminative information (*e.g.*, the glasses region in this case). They propose using a Multiple Instance Learning framework (MIL-SegSVM) that localizes the most discriminative region in each image while learning a classifier to discriminate between classes. We show how our method is also able to estimate the histogram of the discriminative region (*i.e.*, sunglasses) and localize it in the training and test set.

To allow for direct comparison, we used the setup and features of [119]: Our training set is built using images of the first 8 subjects (126 images with sunglasses and 128 without), leaving the remainder for testing (370, equally split among the positive and negative classes).

We represented each image with the BoW model by extracting $10,000$ SIFT features [147] at random scales and positions and quantizing them onto a $1,000$ visual codebook, obtained by performing hierarchical k-means clustering on $100,000$ features randomly selected from the training set. For the first part of the experiment, we compared the results of our classifier to what is obtained using several methods: (1) SVM-Img: a Support Vector Machine (SVM) trained using the entire image, (2) SVM-FS: an SVM trained using a manually labeled discriminative region (in this case, the region of the glasses), (3) MIL-SegSVM: a MIL SVM method proposed by [119]. For MC-1, MC-Pos and MC-Simplex, we proceeded as follows: we built $\mathbf{Z}$ with the label vector and the BoW histograms of each entire image and left the test set labels $\mathbf{Y}^{tst}$ as unknown entries. For the MC-Simplex case, we further preprocessed $\mathbf{Z}$ by dividing each histogram in $\mathbf{Z}_X$ by its sum.

The performance, measured using the area under ROC curve (AUROC), is shown in Table 3.7. These results indicate both the fully supervised (SVM-FS) and the MIL approach (MIL-SegSVM) are more robust to the noise introduced by non-discriminative parts of the images, when compared to training without localization (SVM-Img). However, this is done at either the cost of labeling efforts or by iteratively approximating the solution of the MIL problem, an integer quadratic problem. The matrix completion approaches (MC-1, MC-Pos, MC-Simplex), in turn, are able to surpass these classification scores by solving a convex minimization.

Beyond improving the classification performance, our algorithm is able to localize the discriminative region of interest (the sunglasses region, in this dataset). Recall that the error $\mathbf{E}_\mathbf{X}$ removes the portion of the histogram introduced by the non-discriminative regions of the image. To illustrate this property, after we run the matrix completion classification, we obtain the most discriminative bounding box for all images in the dataset. For each image i in the dataset, we searched for the bounding box that best matches the features of the i-th column of the completed matrix ${\mathbf{z}_\mathbf{X}}^i = \mathbf{h}^i - {\mathbf{e}_\mathbf{X}}^i$ (recall Figure 3.11). We use a

(a) (b)

Figure 3.15: Example images of the CMU-Face dataset. (a) shows the positive class (wearing sunglasses) and (b) shows the negative class (no sunglasses).

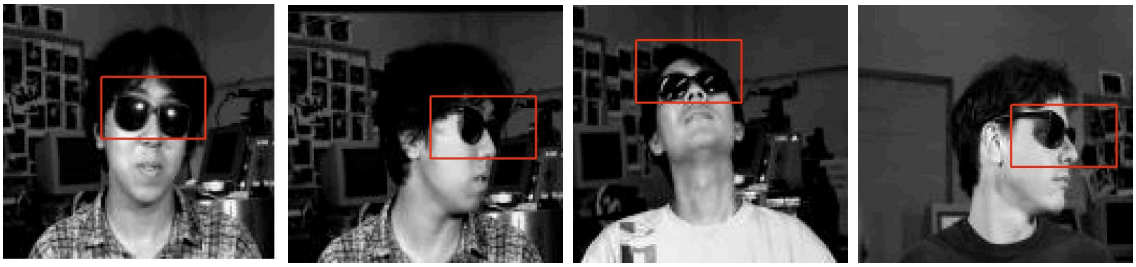


Figure 3.16: A sliding window search shows that histograms corrected by MC-Pos (3.27) are most similar to the discriminative region of the eyes in the images.

sliding window detector varying scale and position using the size criteria in [119] and measure similarity using the χ^2 distance. The results are shown in Figure 3.16 for MC-Pos (similar results were obtained with MC-Simplex). Similarly to MIL-SegSVM, which used a linear SVM score for the subwindow search, our methods correctly localized the eyes region, that discriminates between the classes. Note that MC-1 does not allow to pursue localization of the class representative since it may introduce negative numbers in the histograms.

Table 3.7: AUROC result comparison for the CMU Face dataset.

Method	AUROC
SVM-Img [119]	0.90
SVM-FS [119]	0.94
MIL-SegSVM [119]	0.96
MC-1 [49]	0.96
MC-Pos	**0.97**
MC-Simplex	0.96

Classification in multi-label datasets

In this experiment, we ran our method on two multi-label datasets: MSRC and PASCAL VOC 2007. The MSRC dataset consists of 591 photos distributed among 21 classes, with an average of 3 classes present per image. We mimicked the setup of [125, 126] and used as features histograms of textons [143]. Then, we obtained a 400 word codebook by applying k-means clustering to a random subset of $40,000$ descriptors.

In this task, all training images are labeled with one or several classes, and the goal is to label the test images. Observe that the test image can have several labels (*i.e.*, it's a multi-label classification task). We proceeded as in the experiment described in Sec. 3.2.6. We compared MC-Pos and MC-Simplex with MC-1 and several state-of-the-art multi-label MIL approaches: Multiple Set Kernel MIL (MSK-MIL) by Vijayanarasimhan and Grauman [126], Multi-label Multiple Instance Learning (ML-MIL) by Zha *et al.* [125], Discriminative Multiple Instance Multiple Label model by Yakhnenko and Honavar [39]. We also compared to a one-vs-all linear SVM.

The obtained average AUROC classification scores on the test set using 5-fold cross validation are shown in Table 3.8(a). Results show that our methods outperformed MC-1, thus showing the improvement introduced by the additional constraints and improved loss functions. Moreover, they outperformed results given by state-of-the-art MIL techniques, including the non-linear classifier MSK-MIL. This can be explained by the fact that MIL methods select regions from images to be the positive examples for a class while learning that class boundary. Since possible regions are enumerated by a segmentation algorithm, it is not guaranteed they match exactly the ground truth segmentation. The feature error correction in MC-Pos and MC-Simplex does not require this segmentation step and thus allows for superior results in this weakly supervised multi-label scenario.

We also tested our method in the PASCAL VOC 2007 dataset. This dataset consists

Table 3.8: Classification performance in multi-label datasets.

(a) 5-Fold cross validation average AUROC comparison for image classification tasks on MSRC dataset.

Method	Image
MSK-MIL[126]	0.90
ML-MIL [125]	0.90
DMIML-ℓ_2 [39]	0.91
MC-1 [49]	0.91
MC-Pos	**0.95**
MC-Simplex	0.92
Linear SVM	0.89

(b) Mean Average Precision classification task result comparison in the PASCAL VOC 2007 challenge for two sets of features.

Method	mAP BoW	mAP Overfeat
INRIA_Genetic	0.48	–
MC-1 [49]	0.48	0.73
MC-Pos	0.50	0.73
MC-Simplex	0.50	0.72
Linear SVM	0.49	0.73

of 9963 images labeled with at least one of 20 classes, split into `trainval` and `test` sets. We used the same features as the winning approach (INRIA_Genetic) [117]. This method achieved a mean average precision (mAP) of 0.542. Given that it is a non-linear fusion method, we compare to its simplest feature setting to ensure a fair comparison. We represented each image by extracting dense SIFT features [147] and quantizing them onto a $4,096$ dimension codebook, built by k-means clustering on features randomly selected from the training set followed by ℓ_2 normalization, as implemented in `VlFeat` [148]. INRIA_Genetic reports a mAP of 0.48 for these features. Results in Table 3.8(b) show increased performance for the same features compared to the state of the art circa 2007.Furthermore, we tested using state of the art features obtained from Overfeat, a Convolutional Neural Network trained on ImageNet [111]. We rescaled every image to 221×221 pixels and obtained a single 4096 dimensional feature vector as the output from layer 22 of the network for every image in the dataset. Our independent testing corroborates the results obtained in [149], and the difference to Bag of Words models shows the impressive boost in recognition research in the past 6 years. From these tests, we can conclude Matrix Completion classifiers obtain performances comparable to a linear SVM classifier, while being more versatile in allowing for missing data, as well as noise in labels and features. Moreover, our approach is able to tackle multi-label classification directly and can be useful for object localization, as shown

in the following sections.

Localization in a multi-label dataset

In this section, we propose an alternative exploratory paradigm for the association of labels to regions in the image. The purpose of the method presented herein is not to provide competitive state-of-the-art results for semantic segmentation, but merely to build a working prototype that builds on the histogram representatives naturally obtained by our method, and discuss its advantages and current limitations. Recall that in the two-class example of Sec. 3.2.6, we used each corrected histogram in the training and test set to localize the bounding box containing the most discriminative region. In the multi-label case, however, several classes coexist in one image. Since corrected histograms contain a mixture of classes, they can't be used for class localization in the images.

One possible approach to solve this problem is to pre-segment the test images and use the learned class models to classify each region individually. However, this approach has several drawbacks: 1) having to select a fixed number of segments, 2) the segments are obtained through only texture and color cues, so they might not match the ground truth regions of the classes, and 3) contextual information between segments is lost, which results in poorer classification performance when compared to the classifiers learned on the entire image.

We propose an alternative method that does not suffer from these drawbacks, by explicitly recovering representative histograms for each class. We proceeded as in 3.2.6, but padded the matrix $\mathbf{Z}$ with 21 extra columns where the labels are the identity and the features are unknown, to recover one representative histogram per class (see Figure 3.11(c)). Observe that we do not require segmentation for this classification. For each class in an image (Figure 3.17 (a)), we plot a heatmap of which words belong to the class using its respective histogram (Figure 3.17(b)). Then, we oversegmented each image using the hierarchical segmentation of Arbelaez *et al.* [150] (Figure 3.17(c)). We used code provided by

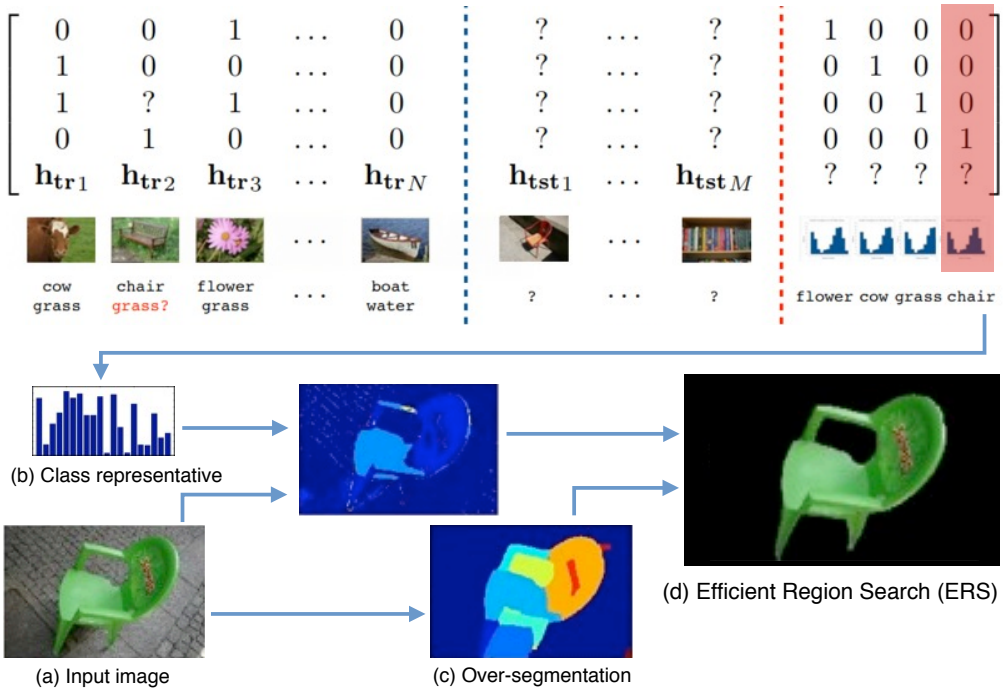


Figure 3.17: Illustration of our method for Matrix Completion localization.

the authors and set the parameter boundary segmentation scale to $k = 0.1$. Last, in order to get the localization for a class in an image, we used the class histograms and the obtained segments for that image as the input to the Efficient Region Search (ERS) method of Vijayanarasimhan and Grauman [151]. ERS selects a group of connected segments (Figure 3.17(d) that maximizes a detection score as measured by an SVM classifier. Since the output of our algorithm is a probability map, we emulated the SVM weight vector by using the class representative subtracted by its mean. We show qualitative results of this approach on Figure 3.2.6 for independent recovery of classes in the same image. The failures of our approach can be generally attributed to one of two cases: class confusion in both the classification and the fact that ERS is applied individually to each class (Figure 3.19(a)); the fact that the solution obtained by ERS is by design a single contiguous region (Figure 3.19(b)).

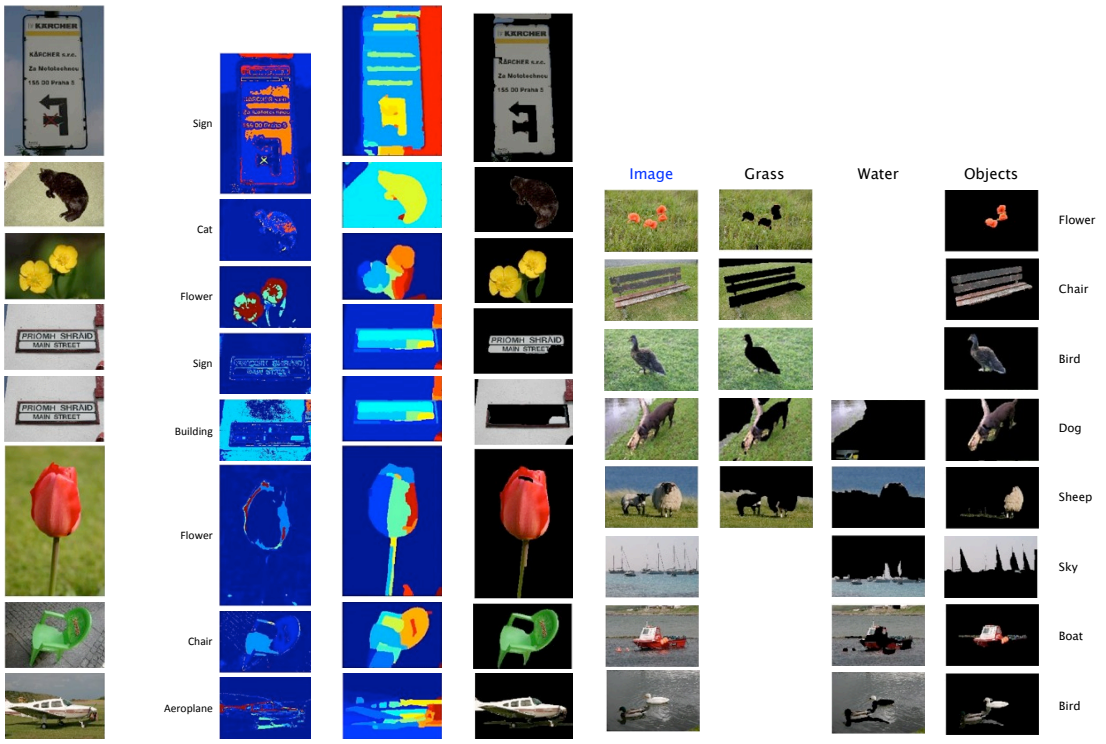
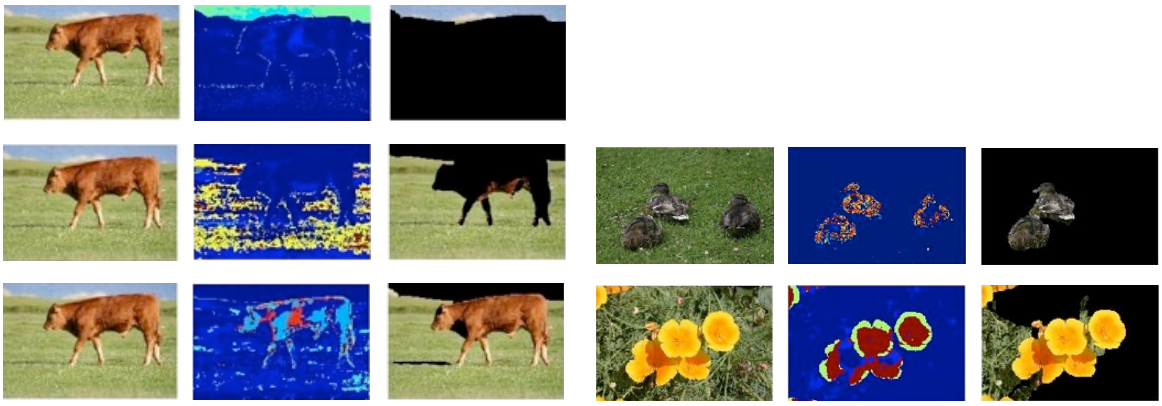


Figure 3.18: Histograms corrected by our method in the MSRC dataset preserve semantic meaning. The input image is shown in (a). The heatmap generated by the class representative histogram is shown in (b). ERS [151] uses the heatmap in (b) and the over segmentation in (c) to produce the segmentation in (d). (e) shows some multi-label segmentation results.



(a) Class confusion and ERS is not multi-label. Top: Sky, Middle: Grass, Bottom: Cow

(b) ERS result is a contiguous region

Figure 3.19: Multi-label segmentation failure cases. Left: Original Image. Middle: Heatmap generated by the class representative histogram. Right: Segmentation obtained by ERS with class representatives.

3.3 Unsupervised learning as a robust PCA problem

In this section, we analyze the application of the unified model proposed in Chapter 2 to Robust PCA, an example of an *unsupervised* learning tasks where label information is not available. We use Robust PCA for the task of background subtraction, which aims to recover a low-rank data matrix $\mathbf{Z}$ from a data matrix $\mathbf{X}$, as

Definition 8 (Robust PCA model)**.**

$$\min_{\mathbf{Z}} \quad \|\mathbf{X}-\mathbf{Z}\|_1 + \lambda\|\mathbf{Z}\|_*, \tag{3.37}$$

As seen in Chapter 2, by using a bilinear factorization of $\mathbf{Z} = \mathbf{U}\mathbf{V}^\top$, the robust PCA model can be equivalently written as (recall (2.6))

$$\min_{\mathbf{U},\mathbf{V}} \quad \|\mathbf{X}-\mathbf{U}\mathbf{V}^\top\|_1 + \frac{\lambda}{2}\left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2\right). \tag{3.38}$$

We show the ALM algorithm proposed in Sec. 2.2 is both faster and more accurate than state-of-the-art nuclear norm algorithms for this problem.

3.3.1 Experimental results

In this section, we validated the lower computational complexity of the algorithm proposed in Sec. 3.3, when the output rank is not known *a priori*. We compared to state-of-the-art nuclear norm and Grassmann manifold methods: GRASTA [59], PRMF [18] and RPCA-IALM [53] in a synthetic and real data experiment for background modeling. We use implementations provided in authors' websites for all baselines. For all experiments, we fix $\mu = 1.05$ initialize $\rho = 10^{-5}$ in Alg. 1. All experiments were run in a desktop with a 2.8 GHz Quad-core CPU and 6 GB RAM.

Table 3.9: Performance comparison of state-of-the-art methods for Robust PCA. Time is in seconds. Error has a factor of 10^{-8}.

Matrix		RPCA-IALM [53]		GRASTA [59]		PRMF [18]		Ours	
N	r	Error	Time	Error	Time	Error	Time	Error	Time
100	3	1.4872	0.3389	226.46	1.7656	3338.7	0.4704	**0.5286**	**0.1734**
200	5	1.5599	2.3575	241.99	2.7282	2687.5	1.0382	**0.7182**	**0.5739**
500	10	3.2595	10.501	263.55	9.5399	1692.4	6.2480	**0.1273**	**3.2373**
1000	15	0.3829	44.111	286.17	23.535	1145.8	30.441	**0.0701**	**14.339**
2000	20	0.6212	196.89	329.11	83.010	808.20	126.95	**0.0308**	**60.658**
5000	25	0.2953	1840.0	379.94	**507.57**	504.08	1307.4	**0.0589**	556.21

Synthetic data We mimicked the setup in [53] and generated low-rank matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. The entries in $\mathbf{U} \in \mathbb{R}^{M\times r}, \mathbf{V} \in \mathbb{R}^{N\times r}$ with $M = N$ and each element sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0, 1)$. Then, we corrupted 10% of the entries with large errors uniformly distributed in the range $[-50, 50]$. The error support was chosen uniformly at random. Like [53], we set $\lambda = \sqrt{N}$ and use the L1 loss. We varied the dimension N and rank r and measured the algorithm accuracies, defined as $\frac{\|\mathbf{Z}-\mathbf{X}\|_2}{\|\mathbf{X}\|_2}$, and the time they took to run. The results in Table 3.9 corroborate experimentally the complexity analysis of the algorithm performed in Sec. 2.2: as N grows significantly larger than r, the smaller runtime complexity of our method allows for equally accurate reconstructions in a fraction of the time taken by RPCA-IALM. While PRMF and GRASTA are also able to outperform RPCA-IALM in time, these methods achieve less accurate reconstructions due to their alternated nature and sampling techniques, respectively.

Real data Next, we compared these methods on a real dataset for background modeling. Here, the goal is to obtain the background model of a slowly moving video sequence. Since the background is common across many frames, the matrix concatenating all frames is a low rank matrix plus a sparse error matrix modeling the dynamic foreground.

We followed the setup of [18] and used the Hall sequence[5]. This dataset consists of 200 frames of video with a resolution of 144×176, and we set the scope of the virtual camera to have the same height, but half the width. We simulated a camera panning

[5]http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

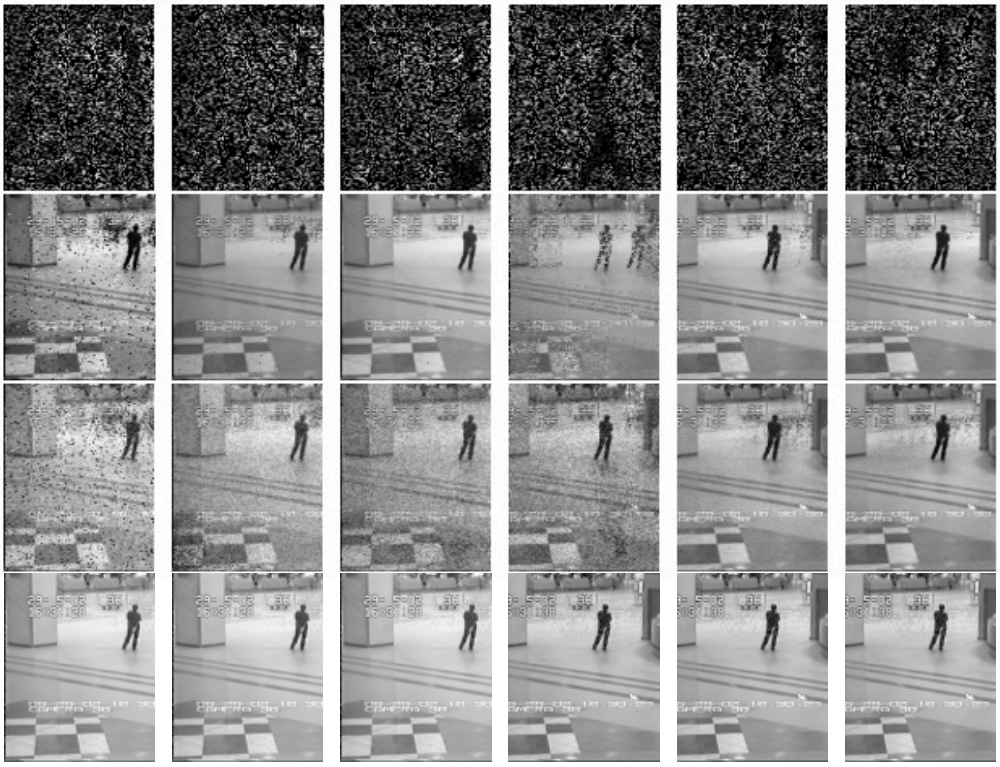


Figure 3.20: Results for background modeling with virtual pan. The first row shows the known entries used for training in frames 40, 70, 100, 130, 170, 200. The remaining rows show the results obtained by PRMF, GRASTA and our method, respectively.

by shifting 20 pixels from left to right in frame 100 to simulate a dynamic background. Additionally, we randomly dropped 70% of the pixels. We proceeded as in the previous synthetic experiment. Fig. 3.20 shows a visual comparison of the reconstruction of several methods. Results corroborate the experiment in Tab. 3.9 and show that the lower accuracies of GRASTA and PRMF yield noisier reconstructions than our method.

Chapter 4

Optimizing hard-rank models when rank is known *a priori*

In Chapter 3, we have focused in soft-rank models and showed that these provide a useful technique where a low-rank solution is sought but its rank is not predetermined. However, many problems in computer vision involve the recovery of shape, appearance or motion representations with a predetermined rank k. Since the recovery of these representations is done from data which is noisy and only partially observed [2, 4, 8], this problem is typically modeled as:

Definition 9 (Rank-k factorization with missing data)**.** Optimization models that aim to recover a rank-k factorization $\mathbf{U}\mathbf{V}^\top \in \mathbb{R}^{M\times N}$ from a data matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$ with missing entries, as

$$\min_{\mathbf{Z}} \quad \|\mathbf{W} \odot (\mathbf{X} - \mathbf{U}\mathbf{V}^\top)\|_F^2, \tag{4.1}$$

where $\mathbf{W} \in \mathbb{R}^{M\times N}$ is a positive weight matrix that can be used to denote missing data (*i.e.*, $w_{ij} = 0$), and the rank-k is specified by the inner dimensions of the product $\mathbf{U}\mathbf{V}^\top$.

Unfortunately, (4.1) is an NP-Hard problem where many state-of-the-art algorithms even fail to reach good local minima [15, 16]. For this reason, the optimization of (4.1) remains an

active research topic, with many work focusing on algorithms that are robust to initialization [3, 8, 17, 18] or initialization strategies [21]. Buchanan *et al.* [19] show that alternated minimization algorithms are subject to flatlining and propose a Newton method to jointly optimize $\mathbf{U}$ and $\mathbf{V}$. Okatani *et al.* [20] show that a Wiberg marginalization strategy on $\mathbf{U}$ or $\mathbf{V}$ is very robust to initialization. However, its high memory usage makes it impractical for medium-size datasets. These methods have also been extended to handle outliers [3, 7, 13]. Ke and Kanade [13] suggest replacing the LS error with the L1 norm, minimized by alternated linear programming. Similarly to the LS case, Eriksson *et al.* [3] show this approach is subject to flatlining and propose a Wiberg extension for L1. Wiberg methods have also been extended to arbitrary loss functions by Strelow [72], but exhibit the same scalability problems as its LS and L1 counterparts. The addition of additional problem specific constraints *e.g.*, orthogonality of $\mathbf{U}$, has also been shown to help algorithms in attaining better minima in structure from motion [8, 17]. However, these methods are not generalizable to several other computer vision problems which are modeled as low-rank factorization problems [8, 9, 10, 11].

In this Chapter, we show that recent soft rank models (recall (2.3)),

$$\min_{\mathbf{Z}} \quad \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 + \lambda \|\mathbf{Z}\|_*, \tag{4.2}$$

are inept to solve problems with a specific predetermined rank-k constraint. To understand why this is the case, let us consider the example of rank-k factorization of a matrix $\mathbf{X}$ under the LS loss with no missing data (*i.e.*, $\mathbf{W} = \mathbf{1}_M \mathbf{1}_N^\top$). For this case, both (4.1) and (4.2) have closed form solutions in terms of the SVD of $\mathbf{X} = \overline{\mathbf{U}}\overline{\Sigma}\overline{\mathbf{V}}^\top$, *i.e.*, $\mathbf{U}\mathbf{V}^\top = \overline{\mathbf{U}}\Sigma_{1:k}\overline{\mathbf{V}}^\top$ and $\mathbf{Z} = \overline{\mathbf{U}}\mathcal{S}_{\frac{\lambda}{2}}(\Sigma)\overline{\mathbf{V}}^\top$. In the case of noisy data, while the former yields the optimal rank-k reconstruction, we need to tune λ in the latter such that $\sigma_{k+1} = 0$. If the λ required to satisfy this constraint is high, it may severely distort the non-zero singular values $\sigma_{1:k}$, resulting in

poor reconstruction accuracy.

Instead, we argue for using our regularized unified model (recall (2.6))

$$\min_{\mathbf{U},\mathbf{V}} \quad \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_F^2 + \frac{\lambda}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 \right). \tag{4.3}$$

While the analysis in Chapter 2 shows that rank restrictions typically lead to local minima when missing data are present, this problem is exacerbated when regularization is not used (*i.e.*, $\lambda = 0$): in addition to gauge freedom[1], it is clear that not all weight matrices $\mathbf{W}$ admit a unique solution [19]. As an extreme example, if $\mathbf{W} = \mathbf{0}$, any choice of $\mathbf{U}$ and $\mathbf{V}$ yields the same (zero) error. Thus, the unregularized factorization in (4.1) will be more prone to local minima than its regularized counterpart (4.3). The two arguments presented against (4.1) and (4.2) provide an argument for choosing our unified model (4.3) and a general guideline for choosing λ: it should be selected as non-zero to ameliorate the local minima problem of (4.1), but small enough such that the first r singular values are not distorted. Moreover, the result in Theorem 1 of Chapter 2 that our model is equivalent to the convex nuclear norm model in (4.2) when k is selected to be big enough allows us to provide a deterministic sequence of initializations for this problem. To summarize, the main contributions of this chapter are:

- In Sec. 4.1, we propose a "rank continuation" deterministic optimization scheme for the NP-Hard factorization problem that avoids local optima in a significant number of cases. This work has been published in [51].
- In Sec. 4.2, we extend the "rank continuation" to the problem of optimizing binary quadratic problems and show an application example for finding correspondences in pairs of images using a max cut formulation. This is currently under review in a journal submission.

[1] for each solution $\mathbf{U}\mathbf{V}^\top$, any solution $(\mathbf{U}\mathbf{R})(\mathbf{R}^{-1}\mathbf{V}^\top)$ where $\mathbf{R} \in \mathbb{R}^{r \times r}$ is an invertible matrix will provide an equal cost.

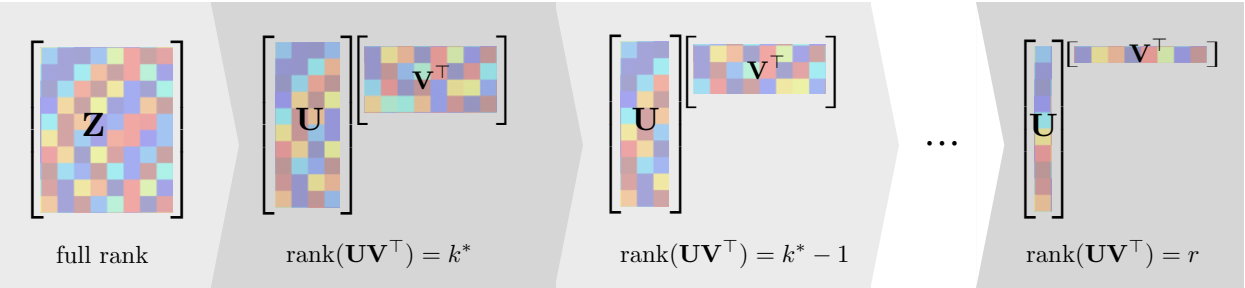


Figure 4.1: Illustration of our proposed rank continuation procedure. Starting with a full rank matrix, we solve the problem and obtain a solution of rank k^*. Under these conditions, Theorem 1 guarantees the our ALM convergens to the global solution and is equivalent to the convex problem. Then we solve a sequence of problems of decreasing rank and initialized with the solution of the previous problem, until the desired rank r is attained.

4.1 Rank continuation for matrix factorization

Given that for any fixed λ, as shown in Theorem 1 of Chapter 2, (4.3) always has a region with no local minima, we propose the following "rank continuation" strategy: we initialize (4.3) with a rank $r \geq k^*$ matrix (*i.e.*, white region of Fig. 2.1, where this problem is equivalent to its convex counterpart), to guarantee its convergence to the global solution. Note that in the absence of an estimate for k^*, we can always use $r = \min(M, N)$. Then, we use this solution as initialization to a new problem (4.3) where the dimensions r of $\mathbf{U}, \mathbf{V}, \mathbf{\Sigma}$ are decreased by one, until the desired rank is attained. This reduction can be done by using an SVD projection. This approach is summarized in Alg. 5 and illustrated in Fig. 4.1. Note this is similar in philosophy to [152] but significantly different in the problem being solved and the continuation path used.

Rank continuation provides a *deterministic* optimization strategy that empirically is shown to find good optima, compared to other baseline algorithms for this family of problems. In particular, we show in the experimental section that global minima of (4.3) are achieved with this strategy in several cases.

Algorithm 5 Rank continuation

Input: $\mathbf{X}, \mathbf{W} \in \mathbb{R}^{M \times N}$, output rank k, parameter λ, an optional estimate of the output rank k^* of (1.3)
  Initialize $\mathbf{U}, \mathbf{V}$ randomly, with $k^* \leq r \leq \min(M, N)$
  Solve for $\mathbf{Z}$ in (4.3) with Alg. 1
  for $r = \text{rank}(\mathbf{Z}) - 1, \ldots, k$ **do**
    SVD: $\mathbf{Z} = \overline{\mathbf{U}}\mathbf{\Sigma}\overline{\mathbf{V}}^\top$
    Rank reduce: $\mathbf{U}_r = \overline{\mathbf{U}}\mathbf{\Sigma}^{\frac{1}{2}}_{1:r}$, $\mathbf{V}_r^\top = \mathbf{\Sigma}^{\frac{1}{2}}_{1:r}\overline{\mathbf{V}}^\top$
    Solve $\mathbf{Z}$ in (4.3) with initialization $\mathbf{U}_r, \mathbf{V}_r$ using Alg. 1
  end for
Output: Complete Matrix $\mathbf{Z}$ with rank k

4.1.1 Experimental results

In this section, we empirically validated the "rank continuation" strategy proposed in Sec. 4.1, in several synthetic and real data problems where the output rank is known *a priori*. We compared our method to state-of-the-art factorization approaches: the damped Newton in [19], the LRSDP formulations in [153] and the LS/L1 Wiberg methods in [3, 20]. Following results reported in the detailed comparisons of [3, 16, 17, 19, 20, 72], we dismissed alternated methods due to their flatlining tendency. To allow direct comparison with published results [17, 19, 20, 153], all methods solved either (4.1) or (4.3) without additional problem specific constraints and we fixed $\lambda = 10^{-3}$. For control, we also compared to two nuclear norm baselines: NN-SVD, obtained by solving (4.2) with the same λ used for other models and projecting to the desired rank with an SVD; NN-λ, obtained by tuning λ in (4.2) so the desired rank is obtained.

Synthetic data

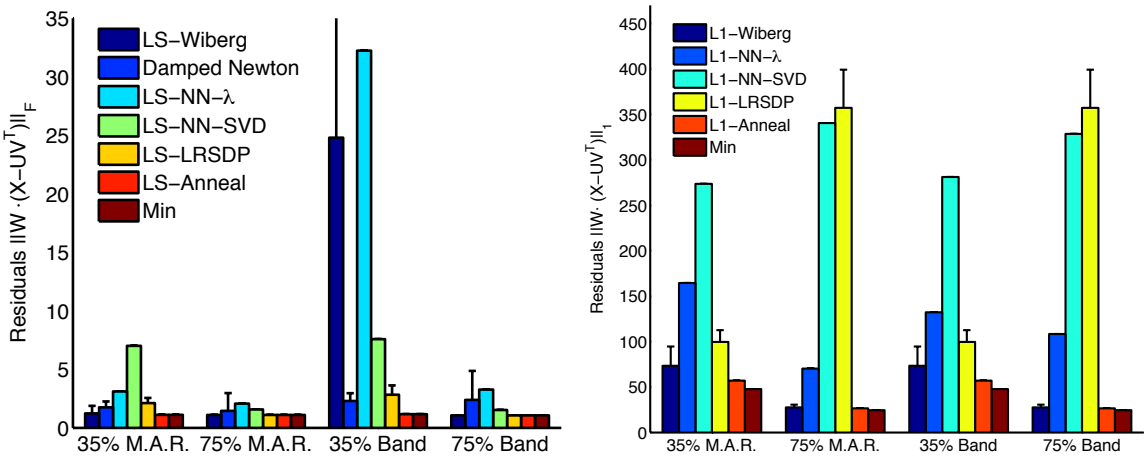
We assessed the convergence performance of our continuation strategy using synthetic data. We performed synthetic comparisons for two loss choices: LS loss $f(\cdot) = \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_F^2$ and the L1 loss $f(\cdot) = \|\mathbf{W} \odot (\mathbf{X} - \mathbf{Z})\|_1$.

For the LS loss, we generated rank-3 matrices $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$. The entries in $\mathbf{U} \in \mathbb{R}^{20\times 3}, \mathbf{V} \in$

$\mathbb{R}^{25\times 3}$ were sampled i.i.d. from a Gaussian distribution $\mathcal{N}(0,1)$ and Gaussian noise $\mathcal{N}(0,0.1)$ was added to every entry of $\mathbf{X}$. For the L1 loss, we proceeded as described for the LS case but additionally corrupted 10% of the entries chosen uniformly at random with outliers uniformly distributed in the range $[-2.5, 2.5]$. We purposely kept the synthetic experiments small, due to the significant memory requirements of the Wiberg algorithms. We varied the percentage of known entries and measured the residual over all *observed* entries, according to the optimized loss function. We chose this measure as it allows for direct comparison between unregularized and regularized models. We ran damped Newton, LRSDP and Wiberg methods 100 times for each test with random initializations.

Fig. 4.2 shows the results for the LS and L1 loss cases. We show two representatives cases for the percentage of known entries (75% and 35%, the breakdown point for L2-Wiberg methods), both for missing data patterns at random (M.A.R.) and with a pattern typical of SfM matrices (Band), generated as in [20]. The theoretical minimum number of entries to reconstruct the matrix is the same as the number of parameters minus factorization ambiguity $Mr + (N - r)(r + 1)$, which for this case is 29.6% [20]. We verified the behavior of all methods when more than 40% of the entries are known is similar to the result shown for 75%.

For the LS case, results in Fig. 4.2(a) show that our deterministic continuation approach always reaches the empirical optima (found as the minimum of all runs of all methods), regardless of the number of known entries or pattern of missing data. Note the minimum error is not zero, due to the variance of the noise. As reported previously [16, 17, 20], we observe that L2-Wiberg is insensitive to initialization for a wide range of missing data entries. However, we note that its breakdown point is not at the theoretical minimum of 35%, due to the lack of regularization. The LRSDP method for optimizing (4.3) outperforms the Wiberg method in this region, suggesting that similar convergence properties of the Wiberg can be obtained without its use of memory. The baseline NN-SVD performed poorly, showing



(a) LS loss (b) L1 loss

Figure 4.2: Comparison of convergence to empirical global minima (Min) for the LS and L1 losses in synthetic data. The minima are found as the minimum of all 100 runs of all methods for each test.

that the estimation of the nuclear norm fits information in its additional degrees of freedom instead of representing it with the true rank.

NN-λ, on the other hand, oversmooths the cost function which might destroy the error function landscape, as is the case for L1 error or when few known entries are available. A visualization of this over-smoothing can be seen in Fig. 4.3, where we reproduce the setup of [15] and plot the landscape of the cost function of (4.3) for a factorization of a 3×3 data matrix (i.e., $\mathbf{W} = [1\ 0\ 1; 0\ 1\ 1; 1\ 1\ 1], \mathbf{X} = [1\ 100\ 2; 100\ 1\ 2; 1\ 1\ 1]$) for $r = 1$ and several values of λ spanning the grey and white areas in Fig. 2.1. From the figure, it can be seen that while convexifying the landscape is appealing for minimization purposes, some global minima might disappear. For the value of λ chosen, however, it seems we benefit from this smoothing without destroying the global minima landscape. This can be seen in Fig. 4.4(a), where we ran Algorithm 1 100 times with random initialization for a 20×25 rank-3 matrix with 50% missing entries generated with band pattern as described in the beginning of this section. For each rank, we plotted the minima attained by the algorithm, and compared it with the path obtained by rank continuation. In this figure, it can be seen

that 1) regularized model exhibits a smaller spread in the number of local minima, and 2) the algorithms directly enforcing the solution converge to several local minima with higher cost, whereas rank continuation attains the correct solution in the final stage. One explanation for why rank continuation attains the optima is that the subspace of rank 3 is contained in the one obtained in the convex problem, i.e., the solution obtained when initializing $\mathbf{Z}$ with full rank. This can be seen in Fig. 4.4(b), where we measured the Normalized Subspace Inclusion (NSI) [154] between the column subspaces $\mathbf{U}_i$ obtained in each rank step i of the continuation (the row subspaces exhibit the same behavior).

$$NSI(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{\text{trace}(\mathbf{U}_i^\top \mathbf{U}_j \mathbf{U}_j^\top \mathbf{U}_i)}{\min(i, j)} \tag{4.4}$$

For the L1 loss case, results in Fig. 4.2(b) show that our continuation strategy no longer attains the empirical optima. We note that this is not surprising since the problem of factorization with missing data is NP-Hard. However, its deterministic result is very close to the optima. Our continuation method regained empirical optimality when only 2% of outliers were present in the data, suggesting a dependency on the noise for the L1 case. In this case, our performance is comparable to what is obtained with the L1-Wiberg algorithm [3] on average. Thus, continuation is a viable alternative to the memory expensive Wiberg method.

Real data

Next, we assessed the results of our continuation approach in real data sequences. We used four popular sequences[2]: a) Dino, for affine SfM; b) Giraffe, for non-rigid SfM, and c) Face and d) Sculpture, both photometric stereo sequences. Their details are summarized in Table 4.1. The dimension of these datasets make the usage of the Wiberg algorithms [3] prohibitive in our modest workstation, due to their memory requirements. For the Sculpture dataset, we treated as missing all pixels with intensity greater than 235 or lower than 20

[2]http://www.robots.ox.ac.uk/~abm/

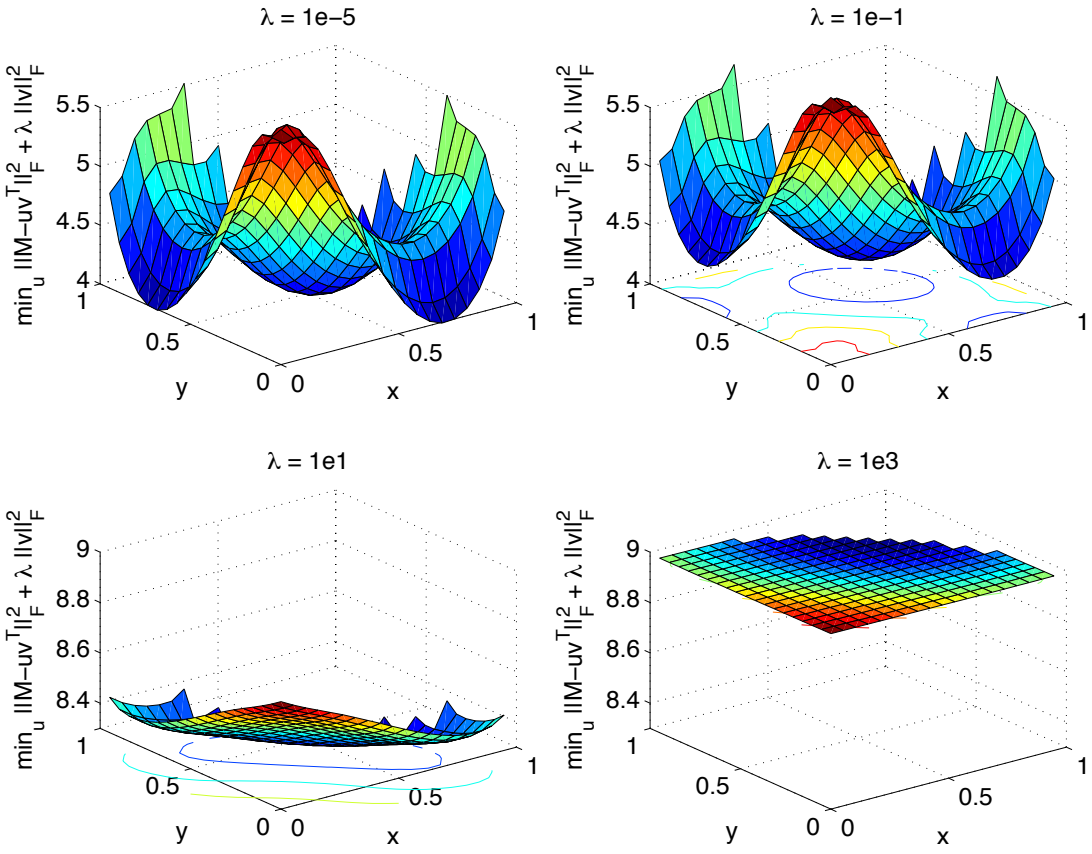
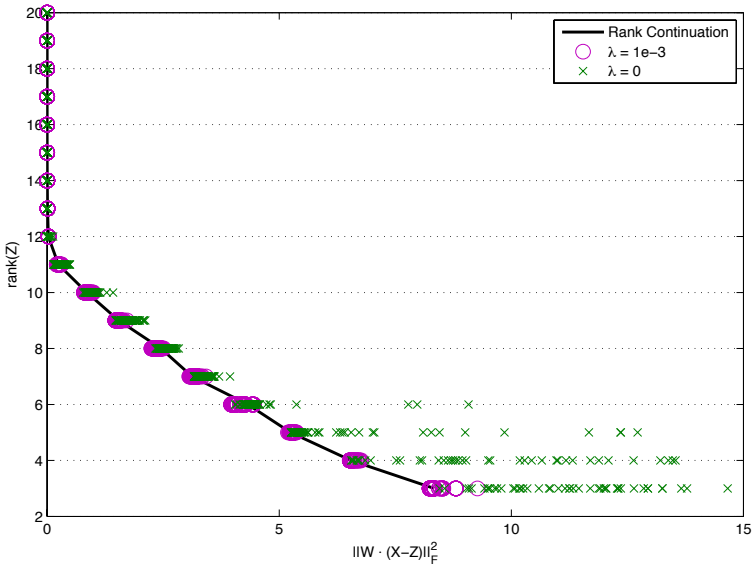
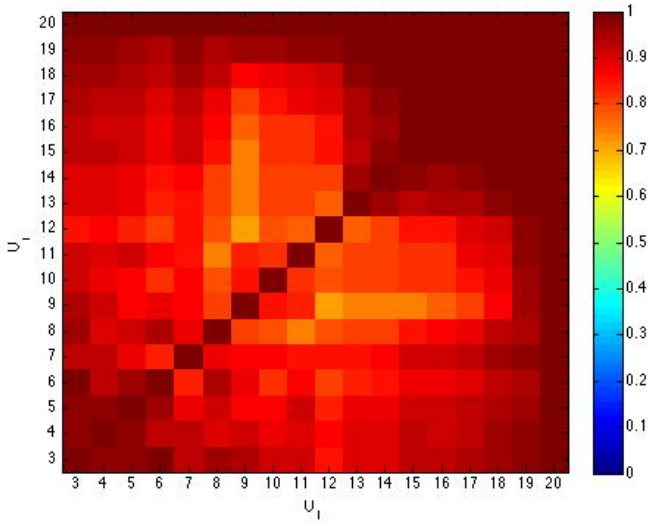


Figure 4.3: Visualization of the cost function in (4.3) for a rank-1 3×3 matrix $\mathbf{X}$ for several values of λ, showing the several local minima existing in the original problem, and that the smoothing induced by the nuclear norm convexifies the problem but its global optima may not necessarily coincide with the original problem's position.



(a)



(b)

Figure 4.4: Intuition behind Rank Continuation. (a) Visualization of minima attained for 100 random runs of Alg. 1 for a rank-3 25×20 matrix $\mathbf{X}$ with and without regularization λ shows the several local minima existing in the cost function landscape, and that the smoothing induced by the nuclear norm allows for avoiding some of these. The rank continuation (solid) attains the global optima in the last rank. (b) Normalized Subspace Inclusion index $NSI(\mathbf{Z}_i, \mathbf{Z}_j)$ measured between the subspaces for each solution step in the continuation for a rank-3 25×20, showing the desired rank-3 subspace is included in the one obtained for the convex region (20).

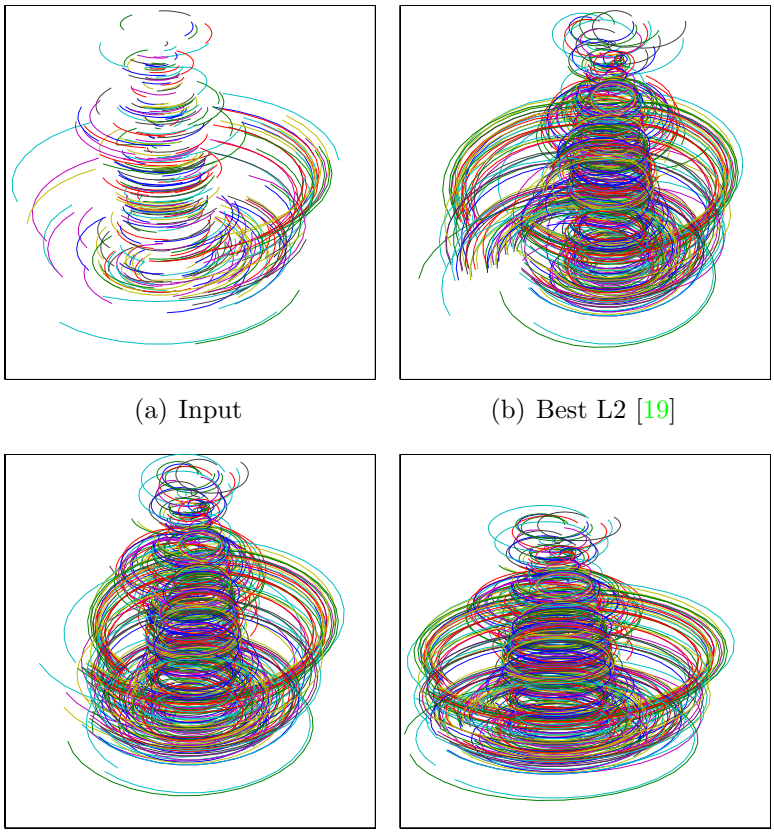
Table 4.1: Real datasets for problems with known output rank-k.

Dataset	Size	Output rank k	Known entries
Dino	319×72	4	28%
Giraffe	240×167	6	70%
Face	2944×20	4	58%
Sculpture	26260×46	3	41%

(*e.g.*, in Fig. 4.6(b), the yellow and purple+black masks, resp.). All other datasets provide $\mathbf{W}$.

Table 4.2 shows a comparison of average error over all *observed* entries for the continuation proposed in Alg. 5 and several methods, according to the loss functions L1/LS. "Best" denotes the best known result in the literature. As explained in Sec. 4.1, we observe that nuclear norm regularized approaches NN-SVD and NN-λ result in bad approximations when a rank restriction is imposed. This can be seen by the high values of λ that have to be used to obtain the desired rank in the variation plots of 4.7. Similar to the results in the synthetic tests, our method always attained or outperformed the state-of-the-art result for the LS loss. The convergence studies in [19, 20] performed optimization on the first three datasets several times with random initializations, so their reported results are suspected by the community to be the global optima for these problems. At the cost of solving several rank constrained problems, our method consistently attains these results in a deterministic fashion, as opposed to state-of-the-art methods which get stuck in local minima several times. As a control experiment, we also ran our continuation strategy for the unregularized case ($\lambda = 0$) on the Dino sequence with LS loss, which resulted in a RMSE of 1.2407. We attribute this to the fact that this case is more prone to local minima, as mentioned in Sec. 4.1.

For the L1 loss, continuation outperforms the state-of-the art in all datasets. It might be argued that problem specific constraints are required to obtain clean reconstructions, but we reiterate the importance of escaping local minima. While there are certainly degenerate scenarios which can only be solved with such constraints [155], Alg. 1 (and consequently, Alg. 5) can be trivially extended to handle such cases. For example, the projection step on



(a) Input (b) Best L2 [19]

(c) RL1NN [17] (d) Continuation

Figure 4.5: Structure from motion Dino sequence. Our L1 continuation method (d) is less prone to local minima and thus can get appealing reconstructions without the use of the additional orthogonality constraints in (c).

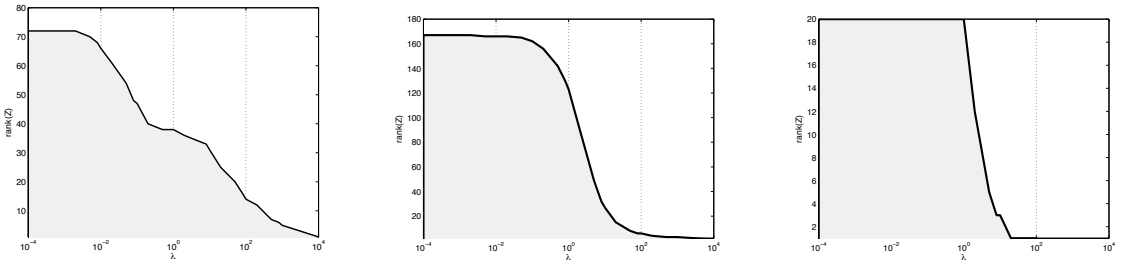
$\mathbf{U}$ for SfM in [8] can be added to Alg. 1 or the problem can be reformulated as a different SDP [153] with a rank constraint, which can be tackled by our continuation strategy in Alg. 5.

4.2 Rank continuation for binary quadratic problems

In this Section, we show that the rank continuation strategy devised in Section 4.1 can be applied as a black box optimization strategy in problems where a rank constraint exists. One special case of rank constrained problems are binary quadratic problems (BQPs), where we

(a) Input (b) Mask (c) NN-λ (d) NNSVD (e) Ours

Figure 4.6: Results for frame 17 of the sculpture sequence for photometric stereo. While (c) smooths out the image and (d) fails to reconstruct it, our continuation approach (e) is able to obtain reconstructions that preserve finer details, such as the imperfections on the cheek or chin.



(a) Dino (b) Giraffe (c) Face

Figure 4.7: Region of equivalence between factorization (4.3) and nuclear norm approaches (4.2) for several real matrix factorization with missing data datasets using the least-squares loss.

Table 4.2: Comparison of LS/L1 average error over all *observed* entries for structure from motion and photometric stereo datasets. State-of-the-art results in best column have been reported in [8, 17, 19, 156]

Dataset	$f(\cdot)$	Best	NN-λ	NN-SVD	Ours
Dino	LS	**1.0847**	6.1699	35.8612	**1.0847**
	L1	0.4283	7.6671	80.0544	**0.2570**
Giraffe	LS	**0.3228**	0.4370	0.6519	**0.3228**
	L1	–	1.8974	11.0196	**0.2266**
Face	LS	**0.0223**	0.0301	0.0301	**0.0223**
	L1	–	0.0287	0.6359	**0.0113**
Sculpt	LS	24.6155	44.5859	31.7713	**22.8686**
	L1	17.753	21.828	33.7546	**12.6697**

wish to recover a binary vector $\mathbf{x}$ that maximizes a pairwise cost given by a matrix $\mathbf{C}$, as

$$\begin{aligned} \max_{\mathbf{x}} \quad & \mathbf{x}^\top \mathbf{C} \mathbf{x} \\ \text{subject to} \quad & \mathbf{x} \in \{-1, 1\} \end{aligned} \tag{4.5}$$

Eq. (4.5) is a special case of a rank constrained model since we can rewrite it without any loss of generality, by lifting the variable product $\mathbf{x}\mathbf{x}^\top$ to a new matrix $\mathbf{X}$, as

$$\begin{aligned} \max_{\mathbf{X}} \quad & \text{trace}(\mathbf{C}^\top \mathbf{X}) \\ \text{subject to} \quad & \text{diag}(\mathbf{X}) = \mathbf{1}, \\ & \mathbf{X} \succeq 0, \\ & \text{rank}(\mathbf{X}) = 1. \end{aligned} \tag{4.6}$$

While important optimization results exist for special cases of these problems — *e.g.*, they become globaly solvable when $\mathbf{C}$ is submodular [157] — the binary constraint in (4.5) makes this general problem NP-Hard and thus intractable in high dimensionality settings. These models are very common in computer vision for problems including partitioning and grouping [158]. In fact, graph cuts [157] and Markov Random Fields [159, 160] approaches are the cornerstone to many computer vision algorithms, such as computing depth fields, regularizing image segmentation, or graphical model inference in object classification.

There are two main approaches for solving general large scale BQPs in computer vision: Semidefinite programming (SDP) approaches, obtained by dropping the rank constraint in (4.6), and spectral approaches[3]. SDP approaches, in particular, work by relaxing intricate constraints of the problem into convex sets in higher-dimensional spaces. Thus, they can been used to obtain upper bounds for combinatorial problems. In fact, these SDP relaxations

[3]as can be seen in [161], the "recipe" for obtaining spectral relaxations is to interpret the binary reformulation in (4.5) as $\|\mathbf{x}\|_2^2 = 1$ and reformulate the problem as a generalized eigenvalue problem.

have been shown to provide better bounds than spectral approaches for many combinatorial problems [158, 162, 163]. Moreover, spectral methods cannot handle inequality constraints, which are necessary in formulations such as segmentation with priors (i.e., biased normalized cuts [164]).

However, three problems remain with SDP approaches. First, they are impaired by the speed of numerical solvers for this problem class. Although off-the-shelf interior point methods can solve SDPs in polynomial time, for many relaxations the exponent in the polynomial complexity bounds is too high for scaling to the large problem sizes typically found in computer vision. Recently, there have been efforts made in the direction of finding scalable and fast approaches for solving this family of SDP problems [165]. However, the effort of obtaining faster algorithmic solutions typically results in bounds that are not as tight as the original SDP formulations. Bie and Cristianini [162] have shown that spectral and SDP relaxations have a continuum of models in between them, and proposed a cascade of relaxations tighter than spectral and looser than SDP. Second, while provable tight bounds have been discovered for specific problems such as the max cut problem, no general result exists on the tightness of bounds when using SDP reformulations for general BQPs. Several efforts have been made in the literature to further tighten the bounds provided by SDP for many problems by adding additional constraints, but often at the cost of exacerbating its scalability problems. Third, the feasible set of the SDP relaxation is convex but not polyhedral, so it is not guaranteed to return a solution in the initial binary domain. Thus, algorithms using these bounds have to rely on postprocessing rounding procedures such as randomized rounding [163], voting schemes or totally unimodular LP projections [166], whose choice varies according to the problem.

Instead of performing the standard SDP relaxation, we note that the low-rank+SDP formulation in (4.6) has a striking similarity to the formulations of (1.2) and the low-rank SDP models of [67]. The surprising results of [67] allow for the feasibility of large scale SDP

problems of this class, by resorting to a factorization model akin to (4.1). Moreover, the deterministic rank continuation strategy we proposed in Sec. 4.1 for the NP-Hard factorization problem that avoids local optima in a significant number of cases is extendable to this family of problems. That is, we propose to solve a sequence of problems that start in a SDP relaxation and gradually decrease the solution rank until they reach a rank-1 problem, which guarantees a binary solution for (4.5). We show experimentally in Sec. 4.2.1 that this continuation strategy avoids local optima in a significant number of cases, akin to the results obtained for the factorization problem described in Sec. 4.1.1. Thus, we believe that rank continuation can be extended to a generic black box optimization strategy for many NP-Hard problems of interest in the computer vision domain that can be formulated as rank constrained problems. Contrary to algorithms designed specifically for each problem, our approach covers graph-optimization problems, unsupervised and supervised classification tasks, and inference on Markov random fields without depending on specific assumptions or problem formulations. For instance, image segmentation using normalized cuts [167], matching using the quadratic assignment problem [168], and solving Markov Random Fields have all been formulated as low-rank SDP problems (cf., [162, 169, 170], respectively).

To exemplify how rank continuation problems can be applied to BQPs, let us consider the graph problem below.

Graph Matching

One BQP problem of interest in computer vision is that of finding correspondences between two images (see Fig. 4.8): in this graph matching problem, each image has a graph of m and n nodes representing interest points, and the goal is to match nodes across images using their similarities and also their shape relationships with neighboring points (modeled as edges on

each graph) [168]. This can be given by the formulation

$$\begin{aligned} \max_{\mathbf{x}\in\{0,1\}^{mn}} \quad & \mathbf{k}^\top \mathbf{x} + \alpha \mathbf{x}^\top \mathbf{K} \mathbf{x} \\ \text{subject to} \quad & \sum_i x_{ij} = 1, \quad \forall 1, \ldots, m \\ & \sum_j x_{ij} \leq 1, \quad \forall 1, \ldots, n \end{aligned} \tag{4.7}$$

which maximizes the point similarities k_{ij} of matched pairs and also the edge similarities $K_{ij,kl}$ between the graphs in both images. In this formulation, the element x_{ij} is 1 if the node i on image 1 is to be matched to node j on image 2 and 0 otherwise. By defining $\hat{\mathbf{K}}$ as

$$\hat{\mathbf{K}} = \begin{bmatrix} 0 & 0.5\mathbf{k}^\top \\ 0.5\mathbf{k} & \alpha\mathbf{K} \end{bmatrix} \tag{4.8}$$

and following the "recipe" mentioned in Sec. 4.2, we lift the binary variable $\mathbf{x}$ to a higher dimensional variable

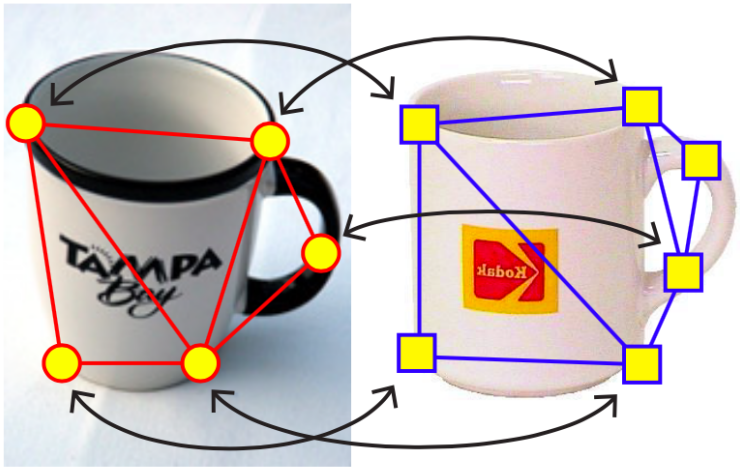


Figure 4.8: Example application for a graph matching BQP: finding correspondences between two figures. Adapted from [168].

$$\hat{\mathbf{X}} = \begin{bmatrix} 1 & \mathbf{x}^\top \\ \mathbf{x} & \mathbf{x}\mathbf{x}^\top \end{bmatrix} \tag{4.9}$$

and rewrite (4.7) as

$$\begin{aligned} \max_{\hat{\mathbf{X}}} \quad & \operatorname{trace}(\hat{\mathbf{K}}\hat{\mathbf{X}}) \\ \text{subject to} \quad & \hat{x}_{1,1} = 1, \\ & 2\operatorname{Diag}(\hat{\mathbf{X}}) = \hat{\mathbf{X}}_{1,:} + \hat{\mathbf{X}}_{:,1}^\top, \\ & \mathbf{H}\operatorname{diag}(\mathbf{X}) = \mathbf{1}_m, \\ & \hat{\mathbf{X}} \succeq 0, \\ & \operatorname{rank}(\hat{\mathbf{X}}) = 1, \end{aligned} \tag{4.10}$$

where $\mathbf{H} = \mathbf{I}_m \otimes \mathbf{1}_n^\top$, and $\hat{\mathbf{X}}_{1,:}$ corresponds to MATLAB notation and denotes the first column of $\hat{\mathbf{X}}$.

The feasible set of the SDP relaxation of (4.10) obtained by dropping the rank constraint is convex but not polyhedral. It contains the set of matrices corresponding to the permutations $\mathbf{x}\mathbf{x}^\top$. But the SDP relaxation solutions discussed above can contain many points not in the affine hull of the constraint set. In particular, it can contain matrices with nonzeros in positions that are zero in the affine hull of the constraint set. So we add additional

constraints corresponding to these zeros, which results in

$$\begin{aligned} \max_{\hat{\mathbf{X}}} \quad & \text{trace}(\hat{\mathbf{K}}\hat{\mathbf{X}}) \\ \text{subject to} \quad & \hat{x}_{1,1} = 1 \\ & 2\text{Diag}(\hat{\mathbf{X}}) = \hat{\mathbf{X}}_{1,:} + \hat{\mathbf{X}}_{:,1}^{\top} \\ & \mathbf{H}\text{diag}(\mathbf{X}) = \mathbf{1}_m \\ & \mathbf{X} \odot \mathbf{M} = \mathbf{0} \\ & \hat{\mathbf{X}} \succeq 0, \\ & \text{rank}(\hat{\mathbf{X}}) = 1, \end{aligned} \tag{4.11}$$

where $\mathbf{M} = \mathbf{I}_m \otimes (\mathbf{1}_n^{\top}\mathbf{1}_n - \mathbf{I}_n) + (\mathbf{1}_m^{\top}\mathbf{1}_m - \mathbf{I}_m) \otimes \mathbf{I}_n$ is the "gangster operator"[4]. The latter rank constraint in (4.11) can be dropped to form an SDP, as

$$\begin{aligned} \max_{\hat{\mathbf{X}}} \quad & \text{trace}(\hat{\mathbf{K}}\hat{\mathbf{X}}) \\ \text{subject to} \quad & \hat{x}_{1,1} = 1 \\ & 2\text{Diag}(\hat{\mathbf{X}}) = \hat{\mathbf{X}}_{1,:} + \hat{\mathbf{X}}_{:,1}^{\top}, \\ & \mathbf{H}\text{diag}(\mathbf{X}) = \mathbf{1}_m, \\ & \mathbf{X} \odot \mathbf{M} = \mathbf{0}, \\ & \hat{\mathbf{X}} \succeq 0. \end{aligned} \tag{4.12}$$

If the optimizer of (4.12) has rank 1, then it is guaranteed to be the optimal result for the original problem (4.7). For the majority of cases, however, the result of (4.12) has higher rank, and thus it is used as an upper bound for (4.7) in the input to a heuristic randomized rounding algorithm [163].

[4] $\mathbf{M}$ is known in the literature as the "gangster operator" since it shoots holes (zeros) in $\mathbf{X}$. We note that additional constraints can be introduced to tighten the bounds obtained by the SDP, as in [67, 165, 171], but this incurs in even more scalability problems as the number of constraints increase.

However, the rank of the resulting SDP has been shown to provide useful information when computing bounds [171] or providing strategies for minimization [152]. In fact, if we examine the equivalent reformulation of (4.11) in the light of the observations in Chapter 2, it is clear that BQPs are a special case of rank constrained problems. Thus, the rank continuation proposed in Sec. 4.1 is directly applicable to this problem class. After reformulating the original BQP as an equivalent low-rank formulation, we can solve it by a sequence of problems starting with a convex problem (4.12) and decreasing the rank until a rank 1 problem is achieved. In Sec. 4.2.1, we show that the rank continuation strategy performs competitvely (and even outperforms in some cases) state-of-the-art algorithms specifically designed for graph matching.

4.2.1 Experimental results

In this section, we compare our rank continuation with SDPCut and several rounding methods, as well as state of the art methods for approximating SDPs in the BQP graph problem of Graph Matching.

Graph Matching

This section reports experimental results on two datasets (one synthetic and one real) and compares our method against the state-of-the-art algorithm for graph matching in computer vision [168]. As a baseline, we also compared to spectral matching [172] and the use of the minimization algorithm with the rank-1 constraint directly imposed in [67] with a random initialization (sdplr1).

This experiment performed a comparative evaluation of four algorithms on randomly synthesized graphs following the experimental protocol of [168]. For each trial, we constructed two identical graphs, G1 and G2, each of which consists of 10 inlier nodes and later we added outlier nodes to both graphs. For each pair of nodes, the edge is randomly generated according to the edge density parameter $\rho \in [0, 1]$. Each edge in the first graph was assigned

a random edge score distributed uniformly and the corresponding edge in the second graph is perturbed by adding a random Gaussian noise $\mathcal{N}(0, \sigma^2)$. The node-affinity was set to zero. We tested the performance of GM methods under three parameter settings. For each setting, we generated 100 different pairs of graphs and evaluated the average accuracy, obtained by comparing the resulting matrix $\mathbf{x}$ with ground truth, and objective ratio w.r.t. to FGM, by computing the cost function in the original model of (4.7) using the obtained $\mathbf{x}$ for each method. In the first setting (Fig. 4.9 left), we increased the number of outliers from 0 to 10 while fixing the noise to zero and considering only fully connected graphs (i.e., $\rho = 1$). In the second case (Fig. 4.9 middle), we perturbed the edge weights by changing the noise parameter σ from 0 to 0.2, while fixing the number of outliers to 0 and $\rho = 1$. In the last case (Fig. 4.9 right), we verified the performance of matching sparse graphs by varying ρ from 1 to 0.3.

Under varying parameters, it can be observed that in most cases, our method achieves state-of-the-art performance in terms of both accuracy and objective ratio, being comparable to FGM [168]. We note that there are cases when FGM achieves higher accuracies. This occurs for results which have a smaller cost function than the one obtained by continuation. This can be attributed to the fact that the optimization problem in (4.7) does not model the rigid matching problem entirely, since rigid motion requires higher order constraints instead of the second order constraints imposed by the model [173].

Additionally, we compared these methods in a real image sequence. The CMU house image sequence is commonly used to test the performance of graph matching algorithms (see(Fig. 4.10)). This dataset consists of 111 frames of a house, each of which has been manually labeled with 30 landmarks. We used Delaunay triangulation to connect the landmarks. The edge weights are computed as the pairwise distance between the connected nodes, as in [168]. We tested the performance of all methods as a function of the separation between frames. We matched all possible image pairs, spaced exactly by $0 : 10 : 90$ frames and com-

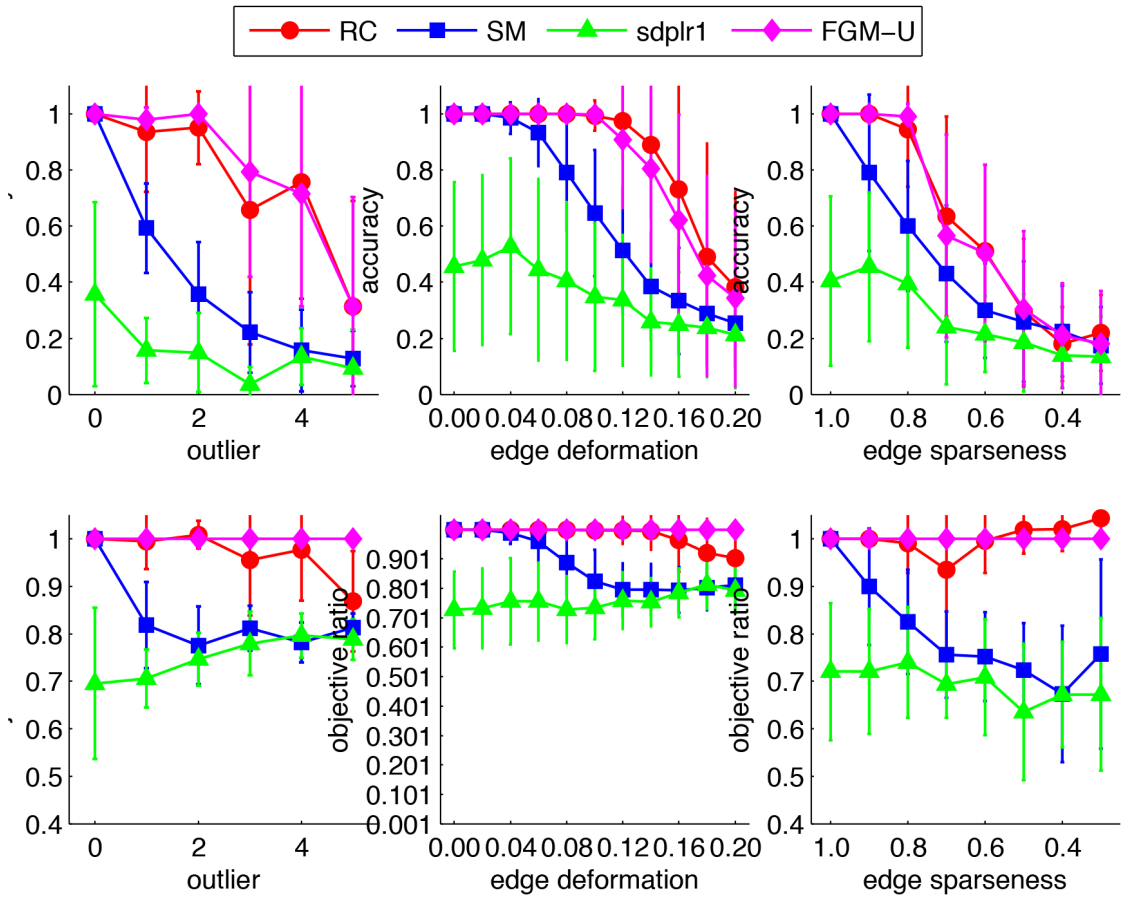
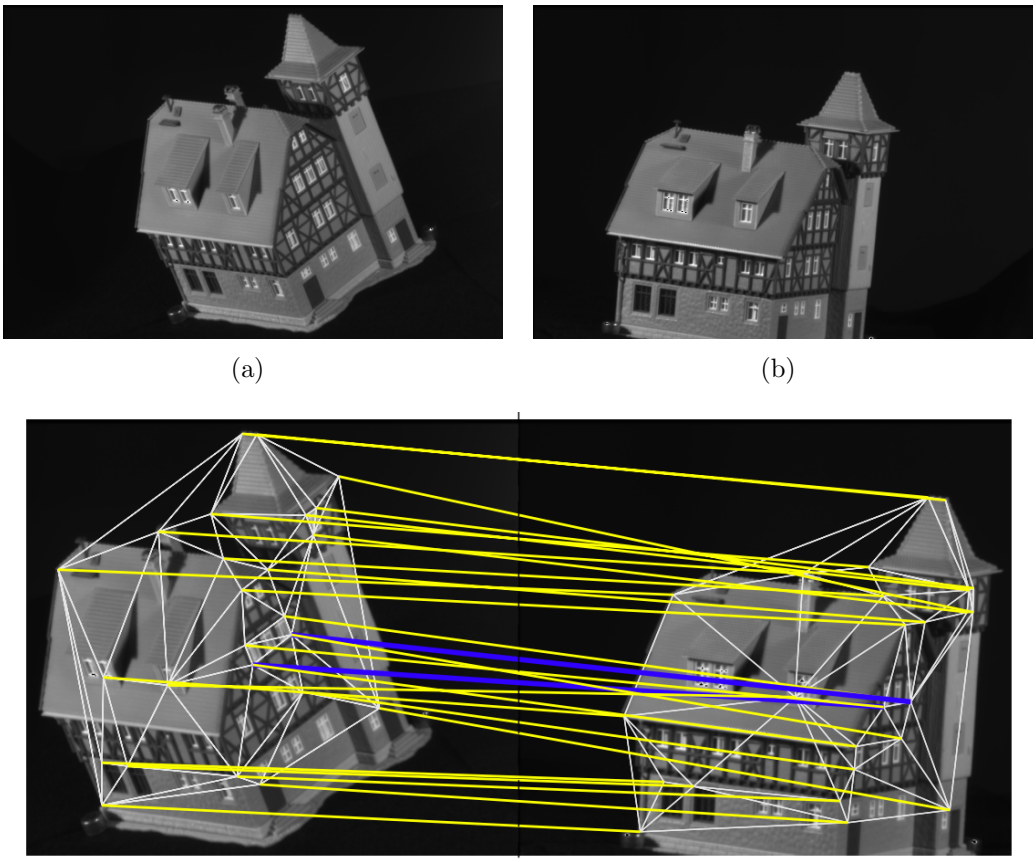


Figure 4.9: Accuracy and objective function result for the graph matching problem of (4.7) in synthetic data for rank continuation (RC), spectral matching (SM), Burer and Monteiro [67] (sdplr1) and Factorized graph matching [168] (FGM). Notice that a ratio bigger than 1 means RC obtains a higher cost function than that of the baseline (FGM). Left: varying number of outliers with no noise and fully connected graphs. Middle: varying edge deformation by changing the noise parameter σ from 0 to 0.2, with zero outliers and fully connected graphs. Right: varying edge sparsity with zero outliers and no noise.



(a) (b)

(c)

Figure 4.10: An example pair of frames (0 and 99) of the House dataset. An example pair of frames with the correspondence generated by FGM [168], where the blue lines indicate incorrect matches.

puted the average matching accuracy and objective ratio per sequence gap. We tested the performance of graph matching methods under two scenarios. In the first case (Fig. 4.11 left) we used all 30 nodes (i.e., landmarks) and in the second one (Fig. 4.11 right) we matched sub-graphs by randomly picking 25 landmarks from each graph. It can be observed that in the first case, FGM, sdplr1 and our method obtained perfect matching of the original graphs. As some nodes became invisible and the graph got corrupted (Fig. 4.11 right), the performance of all the methods degrades. However, our method consistently achieved the best maximum in the objective function.

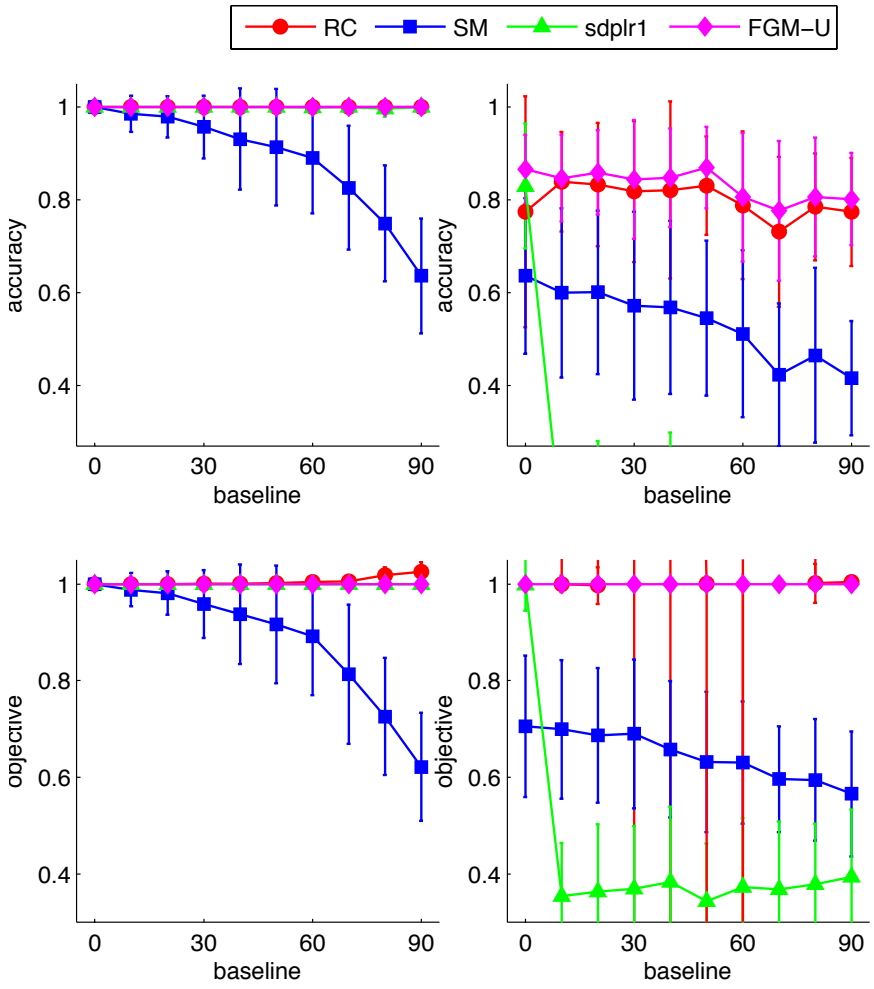


Figure 4.11: Accuracy and objective function value results for the graph matching problem of (4.7) in the CMU House dataset for each baseline (frame distance) using rank continuation (RC), spectral matching (SM), Burer and Monteiro [67] (sdplr1) and Factorized graph matching [168] (FGM). Left: Full data used (30 landmarks). Right: results of randomly picking 25 landmarks on each frame.

The results show that rank continuation outperforms all baselines and performs comparably and in some cases better than FGM, which is considered the state-of-the-art for this problem. We note that this algorithm is sophisticated in its use of the specific structure of the graph matching problem, whereas our strategy is general for rank constrained problems.

Chapter 5

Thesis conclusions and future work

5.1 Major contributions

In our works [44, 51, 62, 63, 64, 174], we answered questions that we summarize below:

Rank continuation, unified model and algorithms The theoretical results and algorithms presented in Chapter 2 show that future work in factorization algorithms should optimize the presented unified model, since it subsumes and inherits benefits of both traditional factorization and the nuclear norm regularized approaches. Based on this analysis, in Chapter 4 we proposed a deterministic "rank continuation" strategy that outperforms state-of-the-art factorization approaches in several computer vision applications with outliers and missing data. Preliminary results show that this strategy is also generalizable for binary quadratic problems such as the quadratic assignment problem. In Chapters 2 and Section 3.2, we have presented Augmented Lagrange Multiplier and Fixed-Point Continuation methods to optimize nuclear norm problems and have studied their convergence properties. An alternative method for incremental nuclear norm optimization, not included in this thesis, can be found in [174].

Robust regression We addressed the problem of robust discriminative learning, and presented a convex formulation for Robust Regression (RR). Our approach jointly learns a

regression, while removing the outliers that are not correlated with labels or regression outputs. We illustrated the benefits of RR in several computer vision problems including facial attribute detection, head pose estimation, and image/video classification. We showed that by removing outliers, our methods consistently learn better representations and outperform state-of-the-art methods in both the linear and kernel spaces (using homogeneous kernel maps). Finally, our approach is general and can be easily applied to robustify other subspace methods such as partial least square or canonical correlation analysis.

Weakly supervised image classification and localization using matrix completion We formulated the weakly-supervised image classification as a low-rank matrix completion problem. Compared to previous work, our proposed framework has three advantages: (1) Unlike existing solutions based on multiple-instance learning methods, our model is convex. (2) Unlike existing discriminative methods, our algorithm is robust to labeling errors, background noise and partial occlusions. (3) Our method can be used for semantic segmentation, despite its weakly-supervised training set, where class locations are unknown. Experimental validation on several datasets showed that our method outperforms state-of-the-art classification algorithms, while effectively capturing each class appearance and allowing for their localization in images.

5.2 Limitations and future work

At the time of writing of this thesis, some questions still remain open. We provide a summary of these below:

Applications of classification with missing data The robust regression and matrix completion methods proposed in Chapter 3 can naturally deal with missing data in the training set. We presented one possible application when merging two different datasets where only a subset of features are common (*e.g.*, color and BW images). However, classifying with missing data has more domains of application. For instance, it may also be applicable in

secure multi-party classification tasks, as one might be interested in learning from a collective dataset of medical data, where each party only releases a subset of descriptors to protect the privacy of its patients' data [175]. Alternatively, it could provide an additional framework for multi-view learning [176].

Can matrix completion localization be extended to multiple exemplars? One caveat of the localization method presented in Sec. 3.2.6 is that only one representative descriptor can be recovered for each class. At the moment, it is unclear for multi-modal class distributions, which direction the algorithm picks or how to extend matrix completion to provide a subspace to represent the class. A partial answer to this is provided in our work [64], where we provide space-time localization of human actions in video by using a union of subspace clustering approach. As an extension of a component analysis technique, this matrix completion classifier should also be kernelized by either the use of homogeneous kernel maps or the results in Sec. 2.3, to couple the feature error correction and the use of non-linear techniques into a single framework.

How do distributed alternatives for matrix factorization with constraints compare to Wiberg? Our work has shown that, for the factorization problem, ALM is a strong contender in terms of attaining global minimum solutions. However, one problem with this framework is its inability to tackle very large scale datasets, such as the ones in [72, 177]. While there has been a surge of research in distributed algorithms for ALM [178] and parallel implementations for matrix completion using stochastic gradient descent [61], further investigation is required to compare these models to recent work in Wiberg algorithms, which look very promising in terms of its applicability to large-scale problems [72].

Why does Rank Continuation work and are there faster alternatives? One explanation for why the rank continuation presented in Chapter 4 attains good optima is that the solution subspace of rank-k is contained in the one obtained in the convex problem, *i.e.*, the

one obtained when initializing $\mathbf{Z}$ with full rank. The original convex solution containing the desired subspace opens the potential for more efficient ways to select the desired subspace from the former, rather than having to run multiple iterations of the algorithm with decreasing rank. One potential solution would be to tackle the problem as a combinatorial problem, similarly to [179], but formulating it as a selection problem from the basis obtained in $\mathbf{Z}$ and potentially exploiting totally unimodularity in order to obtain a solution from convex programming, as done in [180].

Can Rank problems help explain/improve deep neural nets? Recently, there has been a surge of impressive results in the area of object classification, provided by features learned by deep neural networks. However, one problem these approaches currently face is the fact that the neural network topology has to be configured manually. While insights exist on how to perform this task [181], it still mainly is done by a process of trial and error, and due to the large size of these networks [111, 182], at the expense of a significant use of computation power.

We notice that in the past, component analysis techniques have been related to neural networks and this connection has been used to explain the inexistence of local minima in the principal component analysis factorization cost function [12], which enabled for least square based algortihms which enabled the use of this technique in very large datasets [52]. Since early termination been shown to enforce sparsity in the networks and can be seen as a connection to l1-normalization [183] (early stopping is also a known trick in l1-minimization norma algorithms for obtaining sparse solutions) and since linear auto-encoder networks can be shown to be equivalent to a hard-rank matrix factorization model (PCA), we wonder if by modeling auto-encoders as a rank problem and replacing them with soft-rank regularizers could help in automatically discovering good topologies.

Can rank continuation be extended to cardinality problems? In Sec. 4.2, we showed that the strategy devised for matrix factorization with missing data problems can be extended to Binary Quadratic Problems which can be reformulated as SDP problems with a rank constraint. Since the nuclear norm (the sum of the singular values) can be seen as an ℓ_1-norm in the matrix domain, one could potentially extend our findings to LASSO-like problems, comprised by an error function and an ℓ_1-norm regularization together with a cardinality constraint, as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{y}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ \text{subject to} \quad & \text{card}(\mathbf{x}) = k \end{aligned} \tag{5.1}$$

These models are used in dictionary learning [184] and regression problems. Decimation algorithms with cardinality constraints are also especially important in computer vision for the simplification of noisy meshes of regular structures obtained from multi-view stereo pipelines [185], as is the case of buildings in e.g., google street view. One might extend the strategy of rank continuation to cardinality problems of (5.1), by solving initially the convex ℓ_1 problems (dropping the cardinality constraint) and then a sequence of problems with decreasing cardinality constraints. Furthermore, the existing study of parameters in LASSO problems and the connections of this problem to its matrix counterpart in [186] could yield important insights about the parameter λ in nuclear norm regularized problems.

Appendix A

Proof of equivalence between LR-SDP and nuclear norm models

To show this, we first note that (2.6) uses the variational formulation of the nuclear norm in (2.7), and that [61] showed the following result:

Lemma 2. *For any* $\mathbf{Z} \in \mathbb{R}^{M \times N}$*, the following holds: If* $\operatorname{rank}(\mathbf{Z}) = k^* \leq \min(M, N)$*, then the minimum of* (2.7) *is attained at a factor decomposition* $\mathbf{Z} = \mathbf{U}_{M \times k^*} \mathbf{V}_{N \times k^*}^\top$*.*

This result allows us to prove the desired equivalence:

Proof. Applying Lemma 2, we can reduce (2.6) to

$$\begin{aligned} & \min_{\mathbf{U},\mathbf{V}} && f(\mathbf{X} - \mathbf{U}\mathbf{V}^\top) + \lambda \|\mathbf{U}\mathbf{V}^\top\|_* \\ = & \min_{\mathbf{Z}, \operatorname{rank}(\mathbf{Z}) = k^*} && f(\mathbf{X} - \mathbf{Z}) + \lambda \|\mathbf{Z}\|_* \\ = & \min_{\mathbf{Z}} && f(\mathbf{X} - \mathbf{Z}) + \lambda \|\mathbf{Z}\|_*. \end{aligned} \tag{A.1}$$

□

Appendix B

Proof of convergence of MC-1/Pos/Simplex

This appendix proves the convergence of FPC in Alg. 4 by the fact that projections onto Convex sets are non-expansive; thus, the composition of gradient, shrinkage and projection steps is also non-expansive. Since the problem is convex, a unique fixed point exists in its optimal solution.

Lemma 3. *The gradient operator $h(\cdot)$ for (3.24), (3.25), (3.26) is non-expansive for step sizes $\tau \in [0, \min(\frac{4|\Omega_Y|}{\lambda\gamma}, \tau_X|\Omega_X|)]$.*

Proof. These values are obtained from (3.29) by noting the gradient of the Log loss function is Lipschitz continuous with $L = 0.25$ and choosing τ_X such that the χ^2 error, for the Non-Negative Orthant, is Lipschitz continuous with $L = 1$. □

Lemma 4. *Let $p_{\mathcal{C}}(\cdot)$ be a projection operator onto any given convex set $\mathcal{C}$. It follows that $p_{\mathcal{C}}(\cdot)$ is non-expansive and $\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\| = \|\mathbf{Z} - \mathbf{Z}^*\|$ iff $p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) = \mathbf{Z} - \mathbf{Z}^*$.*

Proof. For non-expansiveness, [187, Prop. 3.1.3] states that

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 \leq \langle p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*), \mathbf{Z} - \mathbf{Z}^* \rangle. \tag{B.1}$$

Applying the Cauchy-Schwarz inequality to (B.1) yields

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F \leq \|\mathbf{Z} - \mathbf{Z}^*\|_F. \tag{B.2}$$

For the equivalence part, let us write

$$\begin{aligned} \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 = \\ \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 + \|\mathbf{Z} - \mathbf{Z}^*\|_F^2 \\ -2\langle p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*), \mathbf{Z} - \mathbf{Z}^*\rangle, \end{aligned} \tag{B.3}$$

where the inner product can be bounded by (B.1), yielding

$$\begin{aligned} \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 \leq \|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2 \\ +\|\mathbf{Z} - \mathbf{Z}^*\|_F^2 - 2\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\|_F^2. \end{aligned} \tag{B.4}$$

Since our hypothesis $\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*)\| = \|\mathbf{Z} - \mathbf{Z}^*\|$, (B.4) is

$$\|p_{\mathcal{C}}(\mathbf{Z}) - p_{\mathcal{C}}(\mathbf{Z}^*) - (\mathbf{Z} - \mathbf{Z}^*)\|_F^2 \leq 0, \tag{B.5}$$

from which we conclude an equality is in place. □

Theorem 5. *Let* $\mathbf{Z}^*$ *be an optimal solution to* (3.27) *or* (3.28)*. Then* $\mathbf{Z}$ *is also an optimal solution if*

$$\|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| = \|\mathbf{Z} - \mathbf{Z}^*\|. \tag{B.6}$$

Proof. By non-expansiveness of operators $p_{\mathcal{C}}(\cdot)$, $S_\nu(\cdot)$ and $h(\cdot)$ (Lemma 4 and [55, Lemmas

1,2]), we can write

$$\begin{aligned}\|\mathbf{Z}-\mathbf{Z}^*\| = \|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| \leq \\ \leq \|S_\nu(h(\mathbf{Z})) - S_\nu(h(\mathbf{Z}^*))\| \leq \\ \leq \|h(\mathbf{Z}) - h(\mathbf{Z}^*))\| \leq \|\mathbf{Z}-\mathbf{Z}^*\|,\end{aligned} \tag{B.7}$$

so we conclude the inequalities are equalities. Using the second part of the Lemmas, we get

$$\begin{aligned}p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) = \\ = S_\nu(h(\mathbf{Z}^*)) - S_\nu(h(\mathbf{Z})) = h(\mathbf{Z}^*) - h(\mathbf{Z}) = \mathbf{Z}-\mathbf{Z}^*.\end{aligned}$$

Since $\mathbf{Z}^*$ is optimal, by the projected subgradient method and [55, Corollary 1], we have that

$$p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*))) = \mathbf{Z}^* \implies p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}))) = \mathbf{Z}, \tag{B.8}$$

from which we conclude $\mathbf{Z}$ is an optimal solution to (3.23). □

Theorem 6. *A sequence $\{\mathbf{Z}^k\}$ generated by Alg. 4 converges to $\mathbf{Z}^*$, an optimal solution of (3.27) ((3.28), resp.).*

Proof. We can use the same rationale as in [55, Theorem 4], once we note the non-expansiveness of $p_{\mathcal{C}}(\cdot)$, $S_\nu(\cdot)$ and $h(\cdot)$ ensures the composite operator $p_{\mathcal{C}}(S_\nu(h(\cdot)))$ is also non-expansive. Therefore, the sequence $\{\mathbf{Z}^k\}$ lies in a compact set and must have a limit point, which we define as $\hat{\mathbf{Z}} = \lim_{k\to\infty} \mathbf{Z}^k$. Also, for any solution $\mathbf{Z}^* \in \mathcal{Z}^*$, we have

$$\begin{aligned}\|\mathbf{Z}^{k+1}-\mathbf{Z}^*\| = \|p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^k))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| \leq \\ \leq \|S_\nu(h(\mathbf{Z}^k)) - S_\nu(h(\mathbf{Z}^*))\| \leq \\ \leq \|h(\mathbf{Z}^k) - h(\mathbf{Z}^*))\| \leq \|\mathbf{Z}^k - \mathbf{Z}^*\|,\end{aligned} \tag{B.9}$$

so we conclude the sequence $\{\|\mathbf{Z}^k - \mathbf{Z}^*\|\}$ is monotonically non-increasing and culminates in any limit point $\hat{\mathbf{Z}}$, *i.e.*,

$$\lim_{k\to\infty} \|\mathbf{Z}^k - \mathbf{Z}^*\| = \|\hat{\mathbf{Z}} - \mathbf{Z}^*\|. \tag{B.10}$$

On the other hand, by the continuity of $p_{\mathcal{C}}(S_\nu(h(\cdot)))$, we have that the image of $\hat{\mathbf{Z}}$ is

$$p_{\mathcal{C}}(S_\nu(h(\hat{\mathbf{Z}}))) = \lim_{k\to\infty} p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^k))) = \lim_{k\to\infty} \mathbf{Z}^k = \hat{\mathbf{Z}} \tag{B.11}$$

is also a limit point of $\{\mathbf{Z}^k\}$, yielding

$$\|p_{\mathcal{C}}(S_\nu(h(\hat{\mathbf{Z}}))) - p_{\mathcal{C}}(S_\nu(h(\mathbf{Z}^*)))\| = \|\hat{\mathbf{Z}} - \mathbf{Z}^*\|, \tag{B.12}$$

from which we can recall Theorem 5.

□

Bibliography

[1] D. L. Donoho, "Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality," 2000.

[2] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, pp. 137–154, 1992.

[3] A. Eriksson and A. Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the L1 norm," in *CVPR*, 2010.

[4] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *International Journal of Computer Vision*, vol. 29, no. 3, pp. 159–179, 1998.

[5] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *ICML*, 2007.

[6] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *NIPS*, 2009.

[7] F. De La Torre and M. J. Black, "A Framework for Robust Subspace Learning," *International Journal of Computer Vision*, vol. 54, no. 1-3, pp. 117–142, 2003.

[8] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini, "Bilinear modeling via augmented lagrange multipliers (balm)," *IEEE Transactions on Pattern Analysis and Machine*

Intelligence, vol. 34, no. 8, pp. 1496–1508, 2012.

[9] J. Warrell, P. Torr, and S. Prince, "StyP-Boost : A Bilinear Boosting Algorithm for Style-Parameterized Classifiers," in *BMVC*, 2010.

[10] J. B. Tenenbaum and W. T. Freeman, "Separating Style and Content with Bilinear Models," *Neural Computation*, vol. 1283, pp. 1247–1283, 2000.

[11] H. Pirsiavash, D. Ramanan, and C. Fowlkes, "Bilinear classifiers for visual recognition," in *NIPS*, 2009.

[12] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," *Neural Networks*, vol. 2, no. 1, pp. 53–58, 1989.

[13] Q. Ke and T. Kanade, "Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *CVPR*, 2005.

[14] P. Aguiar, J. Xavier, and M. Stosic, "Spectrally optimal factorization of incomplete matrices," in *CVPR*, 2008.

[15] N. Gillis and F. Glineur, "Low-Rank Matrix Approximation with Weights or Missing Data Is NP-Hard," *SIAM Journal on Matrix Analysis and Applications*, vol. 32, no. 4, 2011.

[16] T. Okatani, T. Yoshida, and K. Deguchi, "Efficient algorithm for low-rank matrix factorization with missing components and performance comparison of latest algorithms," in *ICCV*, 2011.

[17] Y. Zheng, G. Liu, S. Sugimoto, S. Yan, and M. Okutomi, "Practical low-rank matrix approximation under robust l1-norm," in *CVPR*, 2012.

[18] N. Wang, T. Yao, J. Wang, and D. Yeung, "A probabilistic approach to robust matrix factorization," in *ECCV*, 2012.

[19] A. M. Buchanan and A. W. Fitzgibbon, "Damped newton algorithms for matrix fac-

torization with missing data," in *CVPR*, 2005.

[20] T. Okatani and K. Deguchi, "On the Wiberg algorithm for factorization with missing components," *International Journal of Computer Vision*, vol. 72, no. 3, pp. 329–337, 2007.

[21] D. W. Jacobs, "Linear fitting with missing data for structure-from-motion," *CVIU*, vol. 82, pp. 206–2012, 1997.

[22] E. Candès and B. Recht, "Exact low-rank matrix completion via convex optimization," in *Allerton Conference*, 2008.

[23] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings American Control Conference*, 2001.

[24] R. Angst, C. Zach, and M. Pollefeys, "The generalized trace-norm and its application to structure-from-motion problems," in *ICCV*, 2011.

[25] Y. Dai, H. Li, and M. He, "A simple prior-free method for non-rigid structure-from-motion factorization," in *CVPR*, 2012.

[26] ——, "Element-wise factorization for n-view projective reconstruction," in *ECCV*, 2010.

[27] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *ICML*, 2010.

[28] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[29] S. Wang and Z. Zhang, "Colorization by matrix completion," in *AAAI*, 2012.

[30] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse

and low-rank decomposition for linearly correlated images," in *CVPR*, 2010.

[31] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in *CVPR*, 2011.

[32] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *ICCV*, 2011.

[33] Z. Zhang, X. Liang, and Y. Ma, "Unwrapping low-rank textures on generalized cylindrical surfaces," in *ICCV*, 2011.

[34] Z. Zhang, Y. Matsushita, and Y. Ma, "Camera calibration with lens distortion from low-rank textures," in *CVPR*, 2011.

[35] G. Zhu and S. Yan, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *ICMM*, 2010.

[36] K. Min, Z. Zhang, J. Wright, and Y. Ma, "Decomposing Background Topics from Keywords by Principal Component Pursuit," in *CIKM*, 2010.

[37] F. Xiong, O. I. Camps, and M. Sznaier, "Dynamic context for tracking behind occlusions," in *ECCV*, 2012.

[38] A. Argyriou, C. A. Micchelli, and M. Pontil, "On spectral learning," *JMLR*, vol. 11, pp. 935–953, 2010.

[39] O. Yakhnenko and V. Honavar, "Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies," in *BMVC*, 2011.

[40] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-Margin Matrix Factorization," in *NIPS*, 2005.

[41] L. Wolf, H. Jhuang, and T. Hazan, "Modeling Appearances with Low-Rank SVM," in *CVPR*, 2007.

[42] N. Loeff and A. Farhadi, "Scene discovery by matrix factorization," in *ECCV*, 2008.

[43] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face Liveness Detection from A Single Image with Sparse Low Rank Bilinear Discriminative Model," in *ECCV*, 2010.

[44] D. Huang, R. Cabral, and F. De la Torre, "Robust regression," in *ECCV*, 2012.

[45] F. Ojeda, J. a. K. Suykens, and B. De Moor, "Low rank updated LS-SVM classifiers for fast variable selection." *Neural networks : the official journal of the International Neural Network Society*, vol. 21, no. 2-3, pp. 437–49, 2008.

[46] F. R. Bach and I. W. Project-team, "Consistency of Trace Norm Minimization," *JMLR*, vol. 8, pp. 1019–1048, 2008.

[47] D. DeCoste, "Collaborative prediction using ensembles of maximum margin matrix factorizations," in *ICML*, 2006.

[48] R. Tomioka and K. Aihara, "Classifying matrices with a spectral regularization," in *ICML*, 2007.

[49] A. B. Goldberg, X. Zhu, B. Recht, J. ming Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *NIPS*, 2010.

[50] E. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[51] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in *ICCV*, 2013.

[52] F. De la Torre, "A least-squares framework for component analysis," *PAMI*, vol. 34, no. 6, pp. 1041–1055, 2012.

[53] Z. Lin, M. Chen, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," *Mathematical Programming*, 2010.

[54] J.-F. Cai, E. J. Candes, and Z. Shen, "A singular value thresholding algorithm for

matrix completion," *SIAM Journal on Optimization*, vol. 20(4), pp. 1956–1982, 2008.

[55] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[56] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems," *preprint*, 2009.

[57] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theor.*, vol. 56, pp. 2980–2998, June 2010.

[58] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," in *Allerton Conference*, 2010.

[59] J. He, L. Balzano, and A. Szlam, "Incremental gradient on the grassmannian for online foreground and background separation in subsampled video," in *CVPR*, 2012.

[60] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick, "Large-scale image classification with trace-norm regularization," in *CVPR*, 2012.

[61] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization for learning large incomplete matrices," *JMLR*, vol. 99, pp. 2287–2322, 2010.

[62] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Matrix completion for multi-label image classification," in *NIPS*, 2011.

[63] R. Cabral, F. Torre, J. Costeira, and A. Bernardino, "Matrix completion for weakly-supervised multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 121–135, Jan 2015.

[64] E. Adeli Mosabbeb, R. Cabral, F. De la Torre, and M. Fathy, "Multi-label discriminative weakly-supervised human activity recognition and localizatio," in *ACCV*, 2014.

[65] R. S. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear

norm and bilinear factorization approaches for low-rank matrix decomposition," *International Journal of Computer Vision (in review)*, 2015.

[66] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, Aug. 2010.

[67] S. Burer and R. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Mathematical Programming*, vol. 103, no. 3, pp. 427–444, 2005.

[68] R. H. Keshavan and S. Oh, "A gradient descent algorithm on the grassman manifold for matrix completion," *arXiv*, vol. 0910.5260, 2009.

[69] R. Kennedy, L. Balzano, S. J. Wright, and C. J. Taylor, "Online algorithms for factorization-based structure from motion," *CoRR*, vol. abs/1309.6964, 2013.

[70] M. Brand, "Incremental singular value decomposition of uncertain data with missing values," in *IN ECCV*, 2002, pp. 707–720.

[71] R. Liu, Z. Lin, S. Wei, and Z. Su, "Solving principal component pursuit in linear time via l1 filtering," *CoRR*, vol. abs/1108.5359, 2011.

[72] D. Strelow, "General and nested Wiberg minimization," in *CVPR*, 2012.

[73] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[74] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *ECCV*, 2010.

[75] D. Huang, M. Storer, F. De la Torre, and H. Bischof, "Supervised local subspace learning for continuous head pose estimation," in *CVPR*, 2011.

[76] R. Plackett, "Some theorems in least squares," *Biometrika*, vol. 37, no. 1-2, pp. 149–157, 1950.

[77] P. Huber, *Robust Statistics*. Wiley and Sons, 1981.

[78] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley, 2003.

[79] P. Meer, *Robust Techniques for computer vision, Book chapter in Emerging Topics in Computer Vision, G. Medioni and S. Kang (Eds.)*. Prentice Hall, 2004.

[80] P. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[81] J. Gillard, *An Historical Overview of Linear Regression with Errors in both variables*. Cardiff University, School of Mathematics, TR, 2006.

[82] S. V. Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, 1991.

[83] D. Lindley, "Regression lines and the linear functional relationship," *Suppl. J. Roy. Statist. Soc.*, vol. 9, pp. 218–244, 1947.

[84] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[85] P. Torr and A. Zisserman, "Mlesac: A new robust estimator with application to estimating image geometry," *CVIU*, vol. 78, pp. 138–156, 2000.

[86] S. Choi, T. Kim, and W. Yu, "Performance Evaluation of RANSAC Family," in *BMVC*, 2009.

[87] R. Adcock, "A problem in least squares," *Analyst*, vol. 5, no. 2, pp. 53–54, 1878.

[88] C. Kummel, "Reduction of observed equations which contain more than one observed quantity," *Analyst*, vol. 6, pp. 97–105, 1879.

[89] A. Wald, "The fitting of straight lines if both variables are subject to error," *Ann. Math. Statistics*, vol. 11, pp. 285–300, 1940.

[90] J. Gillard and T. Iles, *Method of moments estimation in linear regression with errors in both variables.* Cardiff University, School of Mathematics, TR, 2005.

[91] B. Matei and P. Meer, "Estimation of nonlinear errors-in-variables models for computer vision applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1537–1552, 2006.

[92] C. Chork and P. Rousseeuw, "Integrating a high-breakdown option into discriminant analysis in exploration geochemistry," *Journal of Geochemical Exploration*, vol. 43, pp. 191–203, 1992.

[93] D. Hawkins and G. McLachlan, "High-breakdown linear discriminant analysis," *Journal of the American Statistical Association*, vol. 92, pp. 136–143, 1997.

[94] X. He and W. Fung, "High breakdown estimation for multiple populations with applications to discriminant analysis," *Journal of Multivariate Analysis*, vol. 72, pp. 151–162, 2000.

[95] C. Croux and C. Dehon, "Robust linear discriminant analysis using s-estimators," *Canadian Journal of Statistics*, vol. 29, 2001.

[96] S. Kim, A. Magnani, and S. Boyd, "Robust FDA," in *NIPS*, 2005.

[97] Y. Zhang and D.-Y. Yeung, "Worst-case linear discriminant analysis," in *NIPS*, 2010.

[98] S. Fidler, D. Skocaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 337–350, 2006.

[99] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *CVIU*, vol. 78, no. 1, pp. 99–118, 2000.

[100] H. Jia and A. Martinez, "Support vector machines in face recognition with occlusions,"

in *CVPR*, 2009.

[101] F. De la Torre, "A least-squares framework for component analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1041–1055, 2012.

[102] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[103] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low rank representation," in *NIPS*, 2011.

[104] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, no. 1, pp. 49–67, 2007.

[105] G. Golub and C. V. Loan, "Regression lines and the linear functional relationship," *SIAM Journal on Numerical Analysis*, vol. 17, no. 6, pp. 883–893, 1980.

[106] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "The cmu multi-pose, illumination, and expression (multi-pie) face database," CMU Robotics Institute.TR-07-08, Tech. Rep., 2007.

[107] A. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[108] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[109] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

vol. 27, no. 3, pp. 328–340, 2005.

[110] C. Snoek, M. Worring, J. Gemert, J.-M. Geusebroek, and A. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM MM*, 2006.

[111] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*, 2014.

[112] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 480–492, 2012.

[113] F. Li, G. Lebanon, and C. Sminchisescu, "Chebyshev Approximations to the Histogram χ^2 Kernel," in *CVPR*, 2012.

[114] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar, "Attribute and simile classifiers for face verification," in *ICCV*, Oct 2009.

[115] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *CVPR*, 2013.

[116] B. Marlin, *Missing Data Problems in Machine Learning*. PhD thesis, University of Toronto, 2008.

[117] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[118] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, Oct. 2008.

[119] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother, "Weakly supervised discriminative localization and classification: a joint learning process," in *ICCV*, 2009.

[120] B. Russell, W. Freeman, a.a. Efros, J. Sivic, and a. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections," in *CVPR*, 2006.

[121] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *CVPR*, 2003.

[122] M. Jamieson, A. Fazly, S. Dickinson, S. Stevenson, and S. Wachsmuth, "Learning Structured Appearance Models from Captioned Images of Cluttered Scenes," in *ICCV*, 2007.

[123] C. Yang, M. Dong, and J. Hua, "Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning," in *CVPR*, 2006.

[124] Z.-h. Zhou and M. Zhang, "Multi-instance multi-label learning with application to scene classification," in *NIPS*, 2006.

[125] Z.-j. Zha, X.-s. Hua, T. Mei, J. Wang, and G.-j. Q. Zengfu, "Joint multi-label multi-instance learning for image classification," in *CVPR*, 2008.

[126] S. Vijayanarasimhan and K. Grauman, "What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations," in *CVPR*, 2009.

[127] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2010.

[128] F. Li and C. Sminchisescu, "Convex Multiple Instance Learning by Estimating Likelihood Ratio," in *NIPS*, 2010.

[129] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *CVPR*, 2003.

[130] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *ICCV*, 2001.

[131] C. Desai, D. Ramanan, and C. Fowlkes, "Discriminative models for multi-class object layout," in *ICCV*, 2009.

[132] H. Wang, C. Ding, and H. Huang, "Multi-label linear discriminant analysis," in *ECCV*, 2010.

[133] S. Petridis, W. Liu, and J. Pessiot, "Localizing Objects while Learning Their Appearance," in *ECCV*, 2010.

[134] M. Blaschko and C. Lampert, "Learning to localize objects with structured output regression," *Computer Vision–ECCV 2008*, pp. 2–15, 2008.

[135] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *ECCV*, 2010.

[136] T. Dietterich, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, 1997.

[137] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in *ICML*, 1998.

[138] A. Vezhnevets and J. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *CVPR*, 2010.

[139] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth, "Who's in the Picture?" in *NIPS*, 2004.

[140] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," in *ECCV*, 2006.

[141] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *ICML*, 2008.

[142] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, pp. 2368–2382, 2011.

[143] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *ECCV*, 2006.

[144] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[145] J. Wright, A. Ganesh, S. Rao, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization," in *NIPS*, 2009.

[146] T. Mitchell, *Machine Learning*. McGraw-Hill Education, 1997.

[147] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[148] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[149] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *CVPR*, 2014.

[150] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, May 2011.

[151] S. Vijayanarasimhan and K. Grauman, "Efficient region search for object detection," in *CVPR*, 2011.

[152] B. Mishra, G. Meyer, F. Bach, and R. Sepulchre, "Low-rank optimization with trace

norm penalty," *CoRR*, vol. abs/1112.2318, 2011.

[153] K. Mitra, S. Sheorey, and R. Chellappa, "Large-scale matrix factorization with missing data under additional constraints," in *NIPS*, 2010.

[154] N. P. da Silva and J. P. Costeira, "The normalized subspace inclusion: Robust clustering of motion subspaces." in *ICCV*, 2009.

[155] M. Marques and J. Costeira, "Estimating 3d shape from degenerate sequences with missing data," *CVIU*, vol. 113, no. 2, pp. 261–272, 2009.

[156] L. Wu, A. Ganesh, B. Shi, Y. Matsushita, Y. Wang, and Y. Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *ACCV*, 2010.

[157] Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," in *ICCV*, 1999.

[158] J. Keuchel, C. Schnörr, C. Schellewald, and D. Cremers, "Binary partitioning, perceptual grouping, and restoration with semidefinite programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1364–1379, 2003.

[159] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, June 2008.

[160] J. H. Kappes, B. Andres, F. A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B. X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, and C. Rother, "A comparative study of modern inference techniques for structured discrete energy minimization problems," *CoRR*, vol. abs/1404.0533, 2014.

[161] C. Olsson, A. Eriksson, and F. Kahl, "Solving large scale binary quadratic problems: Spectral methods vs. semidefinite programming," in *CVPR*, 2007.

[162] T. D. Bie and N. Cristianini, "Fast SDP relaxations of graph cut clustering, transduction, and other combinatorial problems," *JMLR*, vol. 7, pp. 1409–1436, Dec. 2006.

[163] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *J. ACM*, vol. 42, no. 6, pp. 1115–1145, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/227683.227684

[164] S. Maji, N. K. Vishnoi, and J. Malik, "Biased normalized cuts," in *CVPR*, 2011.

[165] P. Wang, C. Shen, and A. van den Hengel, "A fast semidefinite approach to solving binary quadratic problems," in *CVPR*, 2013.

[166] J. Maciel and J. P. Costeira, "A global solution to sparse correspondence problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 187–199, 2003.

[167] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[168] F. Zhou and F. De la Torre, "Factorized graph matching," in *CVPR*, 2012.

[169] C. Schellewald and C. Schnörr, "Probabilistic Subgraph Matching Based on Convex Relaxation," in *EMMCVPR*, 2005.

[170] P. H. S. Torr, "Solving markov random fields using semi definite programming," in *AISTATS*, 2003.

[171] J. Peng, H. Mittelmann, and X. Li, "A new relaxation framework for quadratic assignment problems based on matrix splitting," *Mathematical Programming Computation*, vol. 2, no. 1, pp. 59–77, 2010.

[172] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *ICCV*, 2005.

[173] O. Duchenne, F. Bach, I.-S. Kweon, and J. Ponce, "A tensor-based algorithm for high-order graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 12, pp. 2383–2395, Dec 2011.

[174] R. S. Cabral, J. P. Costeira, F. De la Torre, and A. Bernardino, "Fast incremental method for matrix completion: an application to trajectory correction," in *ICIP*, 2011.

[175] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *CVPR*, 2011.

[176] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *CoRR*, vol. abs/1304.5634, 2013.

[177] B. Klingner, D. Martin, and J. Roseborough, "Street view motion-from-structure-from-motion," in *ICCV*, 2013.

[178] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "D-admm: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, May 2013.

[179] F. Jiang, O. Enqvist, and F. Kahl, "A combinatorial approach to l1-matrix factorization," *Journal of Mathematical Imaging and Vision*, 2014.

[180] R. S. Cabral, J. P. Costeira, F. De la Torre, and A. Bernardino, "Optimal no-intersection multi-label binary localization for time series using totally unimodular linear programming," in *ICIP*, 2014.

[181] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *ICCV*, 2009.

[182] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[183] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, pp. 1929–1958, 2014.

[184] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *ICML*, 2009.

[185] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Towards internet-scale multi-view stereo," in *CVPR*, 2010.

[186] E. Grave, G. Obozinski, and F. R. Bach, "Trace lasso: a trace norm regularization for correlated designs," in *NIPS*, 2011.

[187] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, ser. Grundlehren der mathematien Wissenschaften. New York–Heildelberg–Berlin: Springer-Verlag, 2001.