Université Paris Nanterre

Contributions à la Statistique en Grande Dimension

Document de synthèse présenté pour l'obtention d'une

HABILITATION À DIRIGER DES RECHERCHES

Spécialité : **Mathématiques appliquées**

par

Olga Klopp

le 6 décembre 2016

Rapporteurs :

Patrice Bertail	Université Paris Ouest Nanterre la Défense
Emmanuel Candès	Stanford University
Sara van de Geer	ETH Zürich

Devant le jury composé de :

Patrice Bertail	Université Paris Ouest Nanterre la Défense
Gérard Biau	Université Pierre et Marie Curie
Stéphane Boucheron	Université Paris-Diderot
Olivier Guédon	Université Paris-Est Marne La Vallée
Marc Hoffmann	Université Paris-Dauphine
Guillaume Lecué	CNRS et CREST
Alexandre Tsybakov	ENSAE, ParisTech

Remerciements

Tout d'abord, je veux remercier très chaleureusement Alexandre Tsybakov qui, avec une extrême gentillesse et une grande bienveillance, m'a fait découvrir le monde de la statistique non-paramétrique. Il m'a fait confiance, quand il y a six ans, sortie de nulle part, je lui ai demandé de me guider dans cette reconversion scientifique. Je le remercie pour sa disponibilité et le temps qu'il m'a consacré. Sasha m'a proposé un sujet de recherche qui m'a immédiatement passionné, la complétion de matrices. Je lui suis extrêmement reconnaissante pour les nombreuses discussions qui m'ont ouvert des horizons nouveaux et m'ont permis de profiter de la profondeur et de la finesse de sa vision des statistiques et... de la vie. Merci, Sasha, pour toutes ces collaborations scientifiques et d'avoir accepté faire partie du jury de mon HDR !

Je tiens à remercier Sara van de Geer et Emmanuel Candès d'avoir accepté de rapporter sur ce mémoire d'habilitation. Je les remercie pour le temps consacré à cette tâche, l'interêt porté à mon travail et leurs commentaires constructifs et enthousiastes. Avoir comme rapporteurs deux spécialistes de leur envergure est un grand honneur pour moi ! L'une des thématiques principales de ce mémoire est la complétion de matrices, problème introduit dans les travaux fondateurs d'Emmanuel. Qu'il ait accepté de se pencher sur mes contributions à cette thématique est un réel privilège.

J'adresse un merci tout particulier à Patrice Bertail qui m'a guidée dans l'aventure qu'a été la préparation de la soutenance de cette habilitation. Son aide et ses encouragements m'ont été indispensables. Depuis notre première rencontre au CREST, Patrice m'a toujours prodigué conseils et soutien. Grâce à Patrice, au fil des discussions, groupes de travail, mini-cours et exposés j'ai découvert de nouvelles thématiques intéressantes bootstrap, échantillonnage, mathématiques du risque.

Je suis aussi extrêmement heureuse de pouvoir compter dans mon jury Gérard Biau, Stéphane Boucheron, Marc Hoffmann, Guillaume Lecué dont j'admire beaucoup les contributions statistiques ainsi qu' Olivier Guédon, que je n'avais pas le plaisir de connaître personnellement mais dont j'avais lu avec grand intérêt des travaux consacrés aux matrices aléatoires et aux réseaux.

J'ai eu la chance de collaborer avec des co-auteurs exceptionnels. Ces collaborations m'ont énormément enrichie et leurs fruits sont présentés dans ce mémoire. Avec Stéphane Gaïffas, Eric Moulines et Joseph Salmon j'ai eu le plaisir de travailler sur le côté plus appliqué de la complétion de matrices.

Richard Nickl et Karim Lounici m'ont fait bénéficier de leur vastes connaissances. Qu'ils en soient tous remerciés ici. J'adresse un merci tout particulier à Karim pour son aide lors mon séjour au SAMSI et à Richard de m'avoir accueilli au laboratoire de statistiques de Cambridge. C'était un véritable plaisir de collaborer avec Jean Lafond et Matthias Löffler. Un grand merci à Marianna Pensky qui a partagé avec moi la richesse de ses connaissances et sa générosité de caractère. J'ai eu le plaisir de profiter de la ribambelle d'idées d'Alexandra Carpentier et de son intuition mathématique que j'envie ainsi que de nombreuses discussions, pas toutes consacrées aux mathématiques. La profondeur et la vivacité mathématique de Nicolas Verzelen ne cessent de m'étonner. Merci aussi, Nicolas, pour tous ces conseils si pertinents !

J'ai le plaisir de faire partie de deux laboratoires, le CREST et le MODAL'X. Dans ces deux lieux, j'apprécie l'ambiance chaleureuse et très stimulante, la bienveillance de mes collègues, ainsi que l'amitié de certains dentre eux. J'adresse un merci tout particulier aux collègues de Modal'X pour les discussions dans la salle framboise et les délicieux gâteaux ! Je profite de cette occasion pour remercier deux directeurs du MODAL'X, Nathanaël et Antoine, pour leur soutien infaillible !

J'ai gardé pour la fin un remerciement trés spécial à Frédéric sans qui, tout simplement, je ne ferai pas de statistiques. Merci d'avoir cru en moi quand moi je n'y croyais pas, pour ton soutien et tes conseils, de n'être pas toujours d'accord avec moi, du temps passé á parler de mathématiques et de toutes les fautes de français et d'anglais corrigées. Je te le promets : je n'écrirai plus "exposé" avec un e à la fin.

Contents

Chapter 1

Introduction

My first works devoted to the spectral theory of Schrödinger operators are quite far from high dimensional statistics which is the topic of this these. This is the case for [dRT03, Tch05] prepared during my PhD studies at the National Autonomous University of Mexico but also for [dRT07], a paper written once I was recruited as an assistant professor at the same university. These papers reflect the influence of my PhD adviser Rafael del Rio Castillo.

My first years as an assistant professor were also years of scientific questioning. It was during this period that I became interested in statistics having multiples occasions to discuss with colleagues in the department and in the institute of mathematics. The idea of a new scientific project took shape during sabbatical year following my son's birth. I have been very lucky to realize this project thanks to the very valuable support of Alexandre Tsybakov whose human and scientific qualities inspire me. It was during the two post-doctoral years spent with him at CREST that I discovered high-dimensional statistics. It was also Professor Tsybakov who introduced me to the matrix completion problem. I was immediately seduced by this topic. A large part of my papers is devoted to it [Klo11, Klo14, KLMS15, GK17, KLT16, Klo15]. This material is covered in Chapter 2 of this thesis.

In the matrix completion problem one observes only a small number of entries of an unknown matrix. Moreover, the entries that one observes can be perturbed by some noise. From these noisy observations the goal is to recover the unknown matrix. In general, recovery of a matrix from a small number of observed entries is impossible, but, if the unknown matrix has low rank, then accurate and even exact recovery is possible [17, 16, 15].

This problem comes up in many areas including collaborative filtering, multiclass learning in data analysis, system identification in control, global positioning from partial distance information and computer vision, to mention some of them. For example, in NETFLIX recommendation system one observes movie ratings. These ratings form a very large matrix where the rows are users and the columns are movies. Of course each user only rates a small number of movies comparing to the hole NETFIX's offer. The goal of a recommendation system

is to predict missing ratings in order to recommend to each user the movies that he or she might like.

The most popular methods of inference of low-rank matrices are based on minimization of the empirical risk penalized by the nuclear norm with various modifications, see, for example, [17, 33, 48, 51, 38, 45, 29]. In my first paper on this topic [Klo11] I show that in some settings when the observed entries are uniformly distributed it is possible to penalize directly by the rank. An important characteristic of the estimator that I propose in [Klo11] is that it can be computed exactly.

Most of the existing methods of matrix completion rely on the knowledge or a pre-estimation of the standard deviation of the noise. In [GK17] and [Klo14] we propose a new method called $\sqrt{\text{matrix lasso}}$ for approximate low-rank matrix recovery which does not rely on the knowledge or on an estimation of the standard deviation of the noise. This method is inspired by the square-root lasso introduced by Belloni, Chernozhukov and Wang [3] for the vector regression model. We consider two particular settings: matrix completion and matrix regression. In [GK17] jointly with Stéphane Gaïffas we provide empirical results that confirms our theoretical findings and illustrate the fact that using the Frobenius norm instead of the square Frobenius norm as a goodness-of-fit criterion makes the optimal smoothing parameter λ independent of the noise level, allowing for a better stability of the procedure with respect to the noise level, as compared to other state-of-the-art procedures.

Typically, in the matrix completion literature, the sampling scheme is supposed to be uniform. However, in practice, the observed entries are not guaranteed to follow the uniform scheme and its distribution is not known exactly. With this motivation in mind, in [Klo14] I study the restricted nuclear norm penalized estimator under quite general sampling distributions. An important feature of my estimator is that its construction requires only an upper bound on the maximum absolute value of the entries of the unknown matrix. This condition is very mild. A bound on the maximum of the elements is often known in applications. For instance, if the entries of the unknown matrix are some user's ratings it corresponds to the maximal rating. Most of the previous works on matrix completion require more involved conditions on the unknown matrix, for example, the incoherence condition which involves singular vectors of the unknown matrix (see e.g. [14, 37]) or an upper bound on the "spikiness index" (i.e. the quotient of Frobenius and sup norms) as in [45].

In 2012 I was hired as assistant professor at the University Paris Ouest Nanterre la Défense in the Modal'X laboratory. Since my arrival, I have had unfailing support from the laboratory which allowed me to further develop my research. Within Modal'X I continue working on matrix completion problem being especially interested in some questions raised by matrix completion's applications.

First, jointly with Jean Lafond, Éric Moulines and Joseph Salmon [KLMS15], we consider a statistical model where instead of observing a real-valued entry of an unknown matrix we are now able to see only highly quantized outputs. These discrete observations are generated according to a probability distribution

which is parametrized by the corresponding entry of the unknown low-rank matrix. The problem of matrix completion over a finite alphabet has received much less attention than the traditional unquantized matrix completion. This model is well suited for the analysis of voting patterns, preference ratings, or recovery of incomplete survey data, where typical survey responses are of the form "true/false", "yes/no" or "agree/disagree/no opinion" for instance.

Second, jointly with Karim Lounici and Alexandre Tsybakov, in [KLT16] we study robustness to corruptions of matrix completion procedure. It has been shown empirically that uncontrolled and potentially adversarial gross errors that might affect only a few observations are particularly harmful. For example, Xu et al [58] showed that a very popular matrix completion procedure using nuclear norm minimization can fail dramatically even if only a single column has been corrupted. It is particularly relevant in applications to recommendation systems where malicious users try to manipulate the prediction of matrix completion algorithms by introducing spurious perturbations.

A quite popular direction in the matrix completion literature are the thresholding methods which can be divided in two groups: one-step thresholding methods and iterative thresholding methods. Strong theoretical guarantees were obtained for one-step thresholding procedures (see, for example, [38, 20], [Klo11]). Despite these strong theoretical guarantees, these one-step thresholding methods have two important drawbacks: they show poor behavior in practice and only work under the uniform sampling distribution which is not realistic in many practical situations.

Much better practical performances have been shown by iterative thresholding methods as, for example, `SoftImpute` introduced in [44]. These iterative thresholding algorithms are simple to implement, scale to relatively large matrices and achieve competitive errors compared to the state-of-the-art algorithms. In spite of their empirical success, the theoretical guarantees of such iterative thresholding methods are poorly understood. In [Klo15] I provide strong theoretical guarantees, similar to those obtained for nuclear-norm penalization methods and one step thresholding methods for a modification of the `softImpute` algorithm. The results of [Klo15] also answer an important theoretical question: what is the exact minimax rate of convergence for matrix completion problem which has only been known up to a log factor [38].

My meeting with Marianna Pensky leads to two papers on the Varying Coefficient Model [KP13, KP15]. Chapter 3 is dedicated to this topic. The Varying Coefficient Model (VCM) is getting more and more popular in data analysis and has applications in economics, epidemiology, ecology, etc. It provides a more flexible approach than the classical linear regression model and is often used to analyze the data measured repeatedly over time. VCM introduced by Hastie and Tibshirani [34] allows the unknown parameter vector f to depend on the variable t:

$$Y = W^T f(t) + \sigma\xi.$$

Here, $f(\cdot) = (f_1(\cdot), \dots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients. In the applications t represents some characteristics of the

system such as age or time in epidemiological studies.

In [KP13] we propose a novel estimation procedure for f which is based on recent developments in low-rank matrix estimation. To the best of my knowledge, [KP13] is the first non-asymptotic study of VCM. This work lead us to the question of the minimax optimal rates of convergence for this model. We answer this question in [KP15] in the case of sparse high-dimensional Varying Coefficient Model where we introduce a new estimator called "block lasso". This procedure has the advantage of being completely adaptive to sparsity, to heterogeneity of the time dependent covariates and to their possibly spatial inhomogeneous nature.

More recently I have been interested in network models [KTV16]. I present this work in Chapter 4. Networks arise in many areas such as information technology, social life, genetics. They can be studied as graphs, and random graphs analysis has become crucial in order to understand the features of these systems. The main graph integral characteristics are the number of vertices n and the number of edges $|E|$. The relation between these two parameters determines whether a graph is sparse or dense. Most real life networks are sparse and, in general, sparse graphs are more difficult to handle than the dense ones and difficulties increase as the graph get sparser.

In [KTV16] jointly with Alexandre Tsybakov and Nicolas Verzelen we consider a network defined as an undirected simple graph with n nodes. We study the problem of the statistical estimation of the matrix of connection probabilities based on the observations of the adjacency matrix of the network and derive optimal rates of estimation for this problem in the important setting of sparse networks. The results obtained in [KTV16] also yield bounds on the minimax risks for graphon estimation in the L_2 norm when the probability matrix is sampled according to the graphon model. Our results shed light on the differences between estimation under the empirical loss (the probability matrix estimation) and under the integrated loss (the graphon estimation).

The aim of the present memoir is to give an accessible overview of the main results found in the papers described above. Rather than emphasizing technical aspects of the papers we try to explain the results in simple words, put them in context and explain how they relate to the existing literature. All the technical details can be found in the papers available on my website : `http://kloppolga.perso.math.cnrs.fr/publi.html` .

1.1 Notation

We provide a brief summary of the notation. Let A, B be matrices in $\mathbb{R}^{m_1 \times m_2}$.

- For any set I, $|I|$ denotes its cardinal and $\bar{I}$ its complement. Let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.
- For a matrix A, A_{ij} is its (i, j)th entry.

- For two matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$ we define the *scalar product*

$$\langle A, B \rangle = \mathrm{tr}(A^T B).$$

- We denote by $\|A\|_2$ the usual l_2-norm. Additionally, we use the following matrix norms: $\|A\|_*$ is the nuclear norm (the sum of singular values), $\|A\|$ is the operator norm (the largest singular value), $\|A\|_\infty$ is the largest absolute value of the entries:

$$\|A\|_\infty = \max_{i,j} \mid A_{ij} \mid .$$

- $\|A\|_{2,1}$ is the sum of l_2 norms of the columns of A and $\|A\|_{2,\infty}$ is the largest l_2 norm of the columns of A:

$$\|A\|_{2,1} = \sum_{k=1}^{m_2} \|A^k\|_2 \quad \text{and} \quad \|A\|_{2,\infty} = \max_{1 \le k \le m_2} \|A^k\|_2.$$

- The singular value decomposition (SVD) of a matrix A:

$$A = \sum_{j=1}^{\mathrm{rank}\, A} \sigma_j(A) u_j(A) v_j(A)^T, \tag{1.1}$$

  where

  - $\sigma_j(A)$ are the singular values of A indexed in the decreasing order,
  - $u_j(A)$ (resp. $v_j(A)$) are the left (resp. right) singular vectors of A.

- We denote by $S_\lambda(W) \equiv U D_\lambda V'$ the *soft-thresholding* operator where $D_\lambda = \mathrm{diag}\,[(d_1 - \lambda)_+, \ldots, (d_r - \lambda)_+]$, UDV' is the SVD of W, $D = \mathrm{diag}\,[d_1, \ldots, d_r]$ and $t_+ = \max(t, 0)$.

- Let $M = \max(m_1, m_2)$, $m = \min(m_1, m_2)$ and $d = m_1 + m_2$.

- We set $\mathbb{N}_{m_1 \times m_2} = \{(i, j) : 1 \le i \le m_1,\, 1 \le j \le m_2\}$.

- For any vector $\eta \in \mathbb{R}^p$, we denote the standard l_1 and l_2 vector norms by $\|\eta\|_1$ and $\|\eta\|_2$, respectively.

- $\|\cdot\|_{L_2(d\mu)}$ and $\langle \cdot, \cdot \rangle_{L_2(d\mu)}$ are the norm and the scalar product in the space $L_2((0,1), d\mu)$.

- We denote by $\mathbb{R}^{k \times k}_{\mathrm{sym}}$ the class of all symmetric $k \times k$ matrices with real-valued entries.

- The symbol $\lesssim$ means that the inequality holds up to a multiplicative numerical constant and we denote by C positive constant that can vary from line to line.

Chapter 2

Matrix Completion

In recent years, there has been a considerable interest in statistical inference for high-dimensional matrices. One particular problem is matrix completion where one observes only a small number $n \ll m_1 m_2$ of the entries of a high-dimensional $m_1 \times m_2$ matrix A_0 of rank r; it aims at inferring the missing entries. The problem of matrix completion comes up in many areas including collaborative filtering, multi-class learning in data analysis, system identification in control, global positioning from partial distance information and computer vision, to mention some of them. For instance, in computer vision, this problem arises as many pixels may be missing in digital images. In collaborative filtering, one wants to make automatic predictions about the preferences of a user by collecting information from many users. So, we have a data matrix where rows are users and columns are items. For each user, we have a partial list of his preferences. We would like to predict the missing ones in order to be able to recommend items that he may be interested in.

In general, recovery of a matrix from a small number of observed entries is impossible, but, if the unknown matrix has low rank, then accurate and even exact recovery is possible. In the noiseless setting, [17, 16, 15] established the following remarkable result: assuming that it satisfies some low coherence condition, A_0 can be recovered exactly by constrained nuclear norm minimization with high probability from only $n \gtrsim r(m_1 \vee m_2) \log^2(m_1 \vee m_2)$ entries observed uniformly at random.

What makes low-rank matrices special is that they depend on a number of free parameters that is much smaller than the total number of entries. Taking the singular value decomposition of a matrix $A \in \mathbb{R}^{m_1 \times m_2}$ of rank r, it is easy to see that A depends upon $(m_1 + m_2)r - r^2$ free parameters. This number of free parameters gives us a lower bound for the number of observations needed to complete the matrix.

A situation, common in applications, corresponds to the noisy setting in which the few available entries are corrupted by noise. Here we observe a rela-

tively small number of entries of a data matrix

$$Y = A_0 + E$$

where $A_0 = (a_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is the unknown matrix of interest and $E = (\varepsilon_{ij}) \in \mathbb{R}^{m_1 \times m_2}$ is the matrix containing the noise. Noisy matrix completion has been extensively studied recently (e.g., [38, 46, 12, 20]).

2.1 Trace regression and Bernoulli model

Two statistical models have been proposed in the noisy matrix completion literature: the trace regression model (e.g., [38, 46, 12],[Klo14]) and the Bernoulli model (e.g., [20],[Klo15]). In the *trace regression model* we observe couples (X_i, Y_i) satisfying the following relation

$$Y_i = \mathrm{tr}(X_i^T A_0) + \sigma\varepsilon_i,\ i = 1, \dots, n. \tag{2.1}$$

Here $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{m_1 \times m_2}$ are the design matrices, $A_0 \in \mathbb{R}^{m_1 \times m_2}$ is the unknown matrix of interest and ε_i are the noise variables. We assume that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = 1$. Unless stated otherwise, we suppose that the noise variables are independent and *sub-exponential*, i.e., satisfy the following assumption:

Assumption 1. *there exists a positive constant K such that*

$$\max_{i=1,\dots,n} \mathbb{E} \exp\left(|\varepsilon_i|/K\right) < \infty.$$

The trace regression model (2.1) is a quite general model which contains as particular cases a number of interesting problems. For matrix completion, the design matrices X_i are i.i.d copies of a random matrix $X \in \mathbb{R}^{m_1 \times m_2}$ having distribution Π on the set

$$\mathcal{X} = \left\{e_j(m_1)e_k^T(m_2), 1 \le j \le m_1, 1 \le k \le m_2\right\}.$$

Here $e_l(m)$ are the canonical basis vectors in $\mathbb{R}^m$. In this model, Y_i gives us the noisy value of the observed entry and X_i gives its position. Then, the problem of estimating A_0 coincides with the problem of matrix completion with random sampling distribution Π. One of the particular settings of this problem is the Uniform Sampling at Random (USR) matrix completion which corresponds to Π being the uniform distribution .

In the *Bernoulli model* we suppose that each entry $(i, j) \in [m_1] \times [m_2]$ is observed independently of the other entries with probability π_{ij}. Let η_{ij} be the independent Bernoulli variables with parameters π_{ij} and $y_{ij} = \eta_{ij}\left(a_{ij} + \varepsilon_{ij}\right)$. Then, $Y = (y_{ij})$ is the matrix containing our observations. We denote by Ω the random set of observed indices. In a simpler case, the random subset of observed entries is chosen uniformly at random that is, each entry is observed with the same probability p.

Note that in the trace regression model each entry can be sampled multiple times while in the Bernoulli model each entry can be sampled only once. Another difference is that in the trace regression model the number of observations n is fixed while in the Bernoulli model the number of observations $|\Omega|$ is random. In spite of these differences, the results on the minimax optimal estimation obtained for these two models are very similar.

2.2 Rank penalized estimator

The most popular methods of inference of low-rank matrices are based on minimization of the empirical risk penalized by the nuclear norm with various modifications, see, for example, [33, 48, 51, 38, 45, 29]. It is interesting to note that in some settings it is possible to penalize directly by the rank. In [Klo11] I propose a new estimator based on the penalization by the rank and the use of the knowledge of the distribution of the design matrices.

We consider the noisy matrix completion setting. Suppose that we observe n independent random pairs (X_i, Y_i) satisfying the trace regression model (2.1). We assume that the sampling distribution Π is uniform, that is, we suppose that each entry is sampled with the same probability equal to $(m_1 m_2)^{-1}$. In [Klo11] I introduce the following rank-penalized estimator of A_0:

$$\hat{A} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \| A - \mathbf{X} \|_2^2 + \lambda \operatorname{rank}(A) \right\}, \tag{2.2}$$

where

$$\mathbf{X} = \frac{m_1 m_2}{n} \sum_{i=1}^{n} Y_i X_i.$$

The optimization problem (2.2) may equivalently be written as

$$\hat{A} = \arg\min_{k} \left[\min_{A \in \mathbb{R}^{m_1 \times m_2},\ \operatorname{rank}(A)=k} \| A - \mathbf{X} \|_2^2 + \lambda k \right].$$

Here, the inner minimization problem is to compute the restricted rank estimators $\hat{A}_k$ that minimizes the norm $\| A - \mathbf{X} \|_2^2$ over all matrices of rank k. One can write:

$$\hat{A}_k = \sum_{j=1}^{k} \sigma_j(\mathbf{X}) u_j(\mathbf{X}) v_j(\mathbf{X})^T \tag{2.3}$$

where $\mathbf{X} = \sum_{j=1}^{\operatorname{rank} \mathbf{X}} \sigma_j(\mathbf{X}) u_j(\mathbf{X}) v_j(\mathbf{X})^T$ is the singular value decomposition (SVD) of $\mathbf{X}$. Using (2.3), we easily see that $\hat{A}$ has the form

$$\hat{A} = \sum_{j : \sigma_j(\mathbf{X}) \geq \sqrt{\lambda}} \sigma_j(\mathbf{X}) u_j(\mathbf{X}) v_j(\mathbf{X})^T.$$

Thus, the computation of $\hat{A}$ reduces to a simple hard thresholding of singular values in the SVD of $\mathbf{X}$. In [Klo11] I prove the following upper bounds for the estimation error of (2.2) measured in Frobenius and in spectral norms:

Theorem 2. *Assume that* $\|A_0\|_\infty \leq a$ *for some constant* a *and that Assumptions 1 is satisfied. Let* $n > m \log^3(m)$, $\log m \geq 5$ *and* $\sqrt{\lambda} = C(\sigma \vee a)\sqrt{\frac{\log(m)(m_1 \vee m_2)}{n}}$. *Then, with probability at least* $1 - 3/d$, *one has*

(i) $$\frac{\| \hat{A} - A_0 \|_2}{\sqrt{m_1 m_2}} \leq C(\sigma \vee a)\sqrt{\frac{(m_1 \vee m_2)\, r \log(m)}{n}},$$ *and*

(ii) $$\left\|\hat{A} - A_0\right\| \leq C(\sigma \vee a)\sqrt{\frac{m_1 m_2 (m_1 \vee m_2)\log(m)}{n}}.$$

These upper bounds in particular imply that the rank penalized estimator (2.2) is minimax optimal (up to a log factor) both in Frobenius and in spectral norm. This optimality holds for the class of matrices $\mathcal{A}(r, a)$ defined as follows: for given positive r and a, A_0 belongs $\mathcal{A}(r, a)$ if and only if the rank of A_0 is not larger than r and all the entries of A_0 are bounded in absolute value by a. I also prove that $\text{rank}(\hat{A}) \leq \text{rank}(A_0)$.

2.3 Noisy low-rank matrix completion with general sampling distribution

Typically, in the matrix completion literature, the sampling scheme is supposed to be uniform. However, in practice, the observed entries are not guaranteed to follow the uniform scheme and its distribution is not known exactly. For example, in a collaborative filtering setting such as Netflix, where rows of the matrix represent users and columns represent movies, uniform sampling corresponds to assuming all users are equally likely to rate movies and all movies are equally likely to be rated. This assumption is not realistic as some users are much more active than others and some movies are much more popular while others are much less likely to be rated. With this motivation in mind, in [Klo14] I study restricted nuclear norm penalized estimator under general sampling distributions.

Sampling schemes more general than the uniform one were previously considered in [45, 30] where the authors consider penalization using a weighted trace-norm. The weighted trace-norm, used in [45, 30], corrects a specific situation where the standard trace-norm fails. This situation corresponds to a non-uniform distribution where the row/column marginal distribution is such that some columns or rows are sampled with very high probability.

Negahban et al in [45] assumed that the sampling distribution is a product distribution, i.e. the row index and the column index of the observed entries are selected independently. A product distribution assumption does not seem realistic in many cases - e.g. for the Netfix data, it would indicate that all users have the same (conditional) distribution over which movies they rate. An important advantage of the method proposed in [Klo14] is that the sampling distribution does not need to be a product distribution.

Foygel et al in [30] propose a method based on the "smoothing" of the sampling distribution. This procedure may be applied to an arbitrary sampling distribution but requires a priori information on the rank of the unknown matrix. For general sampling distributions the prediction performances of the estimator proposed in [30] are evaluated through bounded l-Lipschitz loss whereas in [Klo14] the estimation error is measured in Frobenius norm. In addition, I show that estimator proposed in [Klo14] is minimax optimal (up to a logarithmic factor).

Assume that we observe n independent random pairs (X_i, Y_i) satisfying the trace regression model (2.1). Let $\pi_{jk} = \mathbb{P}\left(X = e_j(m_1)e_k^T(m_2)\right)$ be the probability to observe the (j,k)-th entry. We denote by $C_k = \sum_{j=1}^{m_1} \pi_{jk}$ the probability to observe an element from the k-th column and by $R_j = \sum_{k=1}^{m_2} \pi_{jk}$ the probability to observe an element from the j-th row. As, unlike [45, 30], in [Klo14] I use the standard trace-norm penalization we need the following assumption on the sampling distribution which guarantees that no row or column is sampled with very high probability:

Assumption 3. *There exists a positive constant $\nu \geq 1$ such that*

$$\max_{i,j} (C_i, R_j) \leq \nu / \min(m_1, m_2).$$

Note that we can easily get an estimation on this upper bound using the empirical frequencies. In order to get bounds in the Frobenius norm, we additionally suppose that each element is sampled with positive probability:

Assumption 4. *There exists a positive constant $\mu \geq 1$ such that*

$$\pi_{jk} \geq (\mu m_1 m_2)^{-1}.$$

In the case of uniform distribution we have that $\nu = \mu = 1$. In [Klo14] I define the following estimator of A_0:

$$\hat{A} = \underset{\|A\|_\infty \leq \mathbf{a}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, A \rangle)^2 + \lambda \|A\|_* \right\}, \tag{2.4}$$

where $\lambda > 0$ is a regularization parameter and $\mathbf{a}$ is an upper bound on $\|A_0\|_\infty$. This is a restricted version of the matrix lasso estimator which is based on a trade-off between fitting the target matrix to the data using least squares and minimizing the nuclear norm.

An important feature of our estimator is that its construction requires only an upper bound on the maximum absolute value of the entries of A_0. This condition is very mild. A bound on the maximum of the elements is often known in applications. For instance, if the entries of A_0 are some users ratings it corresponds to the maximal rating. Most of the previous works on matrix completion

require more involved conditions on the unknown matrix, for example, the incoherence condition (see e.g. [14, 37]) or an upper bound on α_{sp}, the "spikiness index" of the unknown matrix : $\alpha_{sp} = \frac{\sqrt{m_1\, m_2}\, \|A_0\|_\infty}{\|A_0\|_2}$ (see [45]).

The main result of [Klo14] shows the following bound on the normalized Frobenius error of the estimators $\hat{A}$ (2.4):

Theorem 5. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumption 4 and 3. Assume that $\|A_0\|_\infty \leq \mathbf{a}$ for some constant $\mathbf{a}$ and that Assumption 1 holds. Then, with an optimal choice of λ, one has*

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \lesssim (\sigma^2 \vee \mathbf{a}^2)\mu^2\, \nu \frac{\log(d)(m_1 \vee m_2)\mathrm{rank}(A_0)}{n} \tag{2.5}$$

with probability greater than $1 - 3/d$. The symbol $\lesssim$ means that the inequality holds up to a multiplicative numerical constant.

The optimal choice of λ in (2.4) is the following one:

$$\lambda = 3C^*\sigma\sqrt{\frac{2\nu\log(d)}{mn}}. \tag{2.6}$$

where C^* is an absolute numerical constant which depends only on K. If ε_i are $N(0,1)$, then we can take $C^* = 6.5$.

Theorem 5 guarantees, that the prediction error of our estimator is small whenever $n \gtrsim \log(m_1 \vee m_2)(m_1 \vee m_2)\mathrm{rank}(A_0)$. This quantifies the sample size necessary for successful matrix completion. Note that, when $\mathrm{rank}(A_0)$ is small, this is considerably smaller than $m_1 m_2$, the total number of entries. For large m_1, m_2 and small r, this is also quite close to the degree of freedom of a rank r matrix, which is $(m_1 + m_2)r - r^2$.

2.4 High dimensional matrix estimation with unknown variance of the noise

Most of the existing methods of matrix completion rely on the knowledge or a pre-estimation of the standard deviation of the noise. In [GK17] and [Klo14] we consider the problem of high-dimensional matrix estimation from noisy observations with *unknown* variance of the noise. We propose a new method for approximate low-rank matrix recovery which does not rely on the knowledge or on an estimation of the standard deviation of the noise.

Usually, the variance of the noise is involved in the choice of the regularization parameter (see, e.g., (2.6)). The main idea is to use the Frobenius norm instead of the squared Frobenius norm as a goodness-of-fit criterion. Roughly, the idea is that in the KKT condition, the gradient of this square-rooted criterion is the regression score, which is pivotal with respect to the noise level, so that the theoretically optimal smoothing parameter does not depend on the noise

level anymore. This cute idea for dealing with an unknown noise level was first introduced for square-root lasso by Belloni, Chernozhukov and Wang [3] for the vector regression model. We consider two particular settings: matrix completion and matrix regression.

2.4.1 Square-root estimator for Matrix Completion

In [Klo14] I propose a new estimator for the matrix completion problem in the case when the variance of the noise σ is unknown. Assume that we observe n independent random pairs (X_i, Y_i) satisfying the trace regression model (2.1) and define the following estimator of A_0:

$$\hat{A}_{SQ} = \underset{\|A\|_\infty \le \mathbf{a}}{\arg\min} \left\{ \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \langle X_i, A\rangle\right)^2} + \lambda \|A\|_* \right\}, \tag{2.7}$$

where $\lambda > 0$ is a regularization parameter and $\mathbf{a}$ is an upper bound on $\|A_0\|_\infty$. Note that the first term of this estimator is the square root of the data-dependent term of the restricted matrix lasso estimator (2.4). This is similar to the principle used to define the square-root lasso estimator for the usual vector regression model. We consider the case of sub-Gaussian noise:

Assumption 6. *There exists a constant K such that*

$$\mathbb{E}\left[\exp(t\varepsilon_i)\right] \le \exp\left(t^2/2K\right)$$

for all $t > 0$.

Note that condition $\mathbb{E}\varepsilon_i^2 = 1$ implies that $K \le 1$. We can take λ in (2.7) in the following way:

$$\lambda = 6C^* \sqrt{\frac{2\nu \log(m_1 + m_2)}{(m_1 \wedge m_2)n}} \tag{2.8}$$

where C^* is an absolute numerical constant and ν is defined in Assumption 3. If ε_i are $N(0,1)$, then we can take $C^* = 6.5$. Here λ *does not depend* on σ. We consider n large enough, more precisely, n such that

$$n \gtrsim \mu\nu(m_1 \vee m_2)\mathrm{rank}(A_0)\log(d) \tag{2.9}$$

and obtain the following theorem:

Theorem 7. *Let X_i be i.i.d. with distribution Π on $\mathcal{X}$ which satisfies Assumption 3 and 4. Assume that $\|A_0\|_\infty \le \mathbf{a}$ for some constant $\mathbf{a}$ and that Assumption 6 holds. Consider a regularization parameter λ satisfying (2.8) and n satisfying (2.9). Then, with probability greater than $1 - 3/d - 2\exp\{-c_3 n\}$,*

$$\frac{\|\hat{A}_{SQ} - A_0\|_2^2}{m_1 m_2} \lesssim (\sigma^2 \vee \mathbf{a}^2)\mu^2\nu \frac{\log(d)\mathrm{rank}(A_0)(m_1 \vee m_2)}{n}. \tag{2.10}$$

The symbol $\lesssim$ means that the inequality holds up to a multiplicative numerical constant.

Note that condition (2.9) is not restrictive: indeed the sampling sizes n satisfying condition (2.9) are of the same order of magnitude as those for which the normalized Frobenius error of our estimator is small. Thus, Theorem 7 shows that $\hat{A}_{SQ}$ has the same prediction performances as previously proposed estimators which rely on the knowledge of the standard deviation of the noise. In particular $\hat{A}_{SQ}$ is minimax optimal up to a logarithmic factor. This optimality holds for the class of matrices $\mathcal{A}(r, a)$ defined as follows: for given r and a, A_0 belongs to $\mathcal{A}(r, a)$ if and only if the rank of A_0 is not larger than r and all the entries of A_0 are bounded in absolute value by a. Note also that the lower bound obtained in [38] contains the minimum of σ^2 and a^2 whereas in (2.10) we have the maximum, so, in a particular setting when $\sigma \wedge a \to 0$, the dependency on these two parameters given by (2.10) probably is not optimal.

2.4.2 Square-root estimator for Matrix Regression

In [GK17], jointly with Stéphane Gaïffas, we apply the idea of square root estimator to the matrix regression. The *matrix regression model* is given by

$$U_i = V_i A_0 + E_i \qquad i = 1, \dots, n$$

where U_i are $1 \times m_2$ vectors of response variables, V_i are $1 \times m_1$ vectors of predictors, A_0 is an unknown $m_1 \times m_2$ matrix of regression coefficients of rank r and E_i are random $1 \times m_2$ vectors of noise with independent entries and mean zero. Set $V = \left(V_1^T, \dots, V_n^T\right)^T$, $U = \left(U_1^T, \dots, U_n^T\right)^T$.

In [GK17] we propose a new square-root type estimator of A_0:

$$\hat{A} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \|U - V A\|_2 + \lambda \|VA\|_* \right\},$$

where $\lambda > 0$ is a regularization parameter. This estimator can be formulated as a solution to a conic programming problem and we prove the following upper bound on the estimation error of $\hat{A}$:

Theorem 8. *Assume that ε_{ij} are independent $N(0, 1)$ and (2.12) is satisfied. Then, for an optimal choice of λ, we have that*

$$\left\| V \left(\hat{A} - A_0 \right) \right\|_2^2 \lesssim \sigma^2 (m_2 + v) \operatorname{rank}(V A_0)$$

with probability at least $1 - 2\exp\{-c(m_2 + v)\}$. Here $v = \operatorname{rank}(V)$.

To the best of my knowledge, it is an interesting open problem whether or not the upper bound given by Theorem 8 is optimal. Previously matrix regression with unknown noise variance was considered in [8, 32]. These two papers study rank-penalized estimators. Bunea et al [8] proposed an unbiased estimator of σ which required an assumption on the dimensions of the problem. This assumption excludes an interesting case when the sample size is smaller than the number of covariates; our method can be applied to this case.

The method proposed in [32] can be applied to this last case under the following condition on the rank of the unknown matrix A_0:

$$\operatorname{rank}(A_0) \leq \frac{C_1(n\,m_2 - 1)}{C_2\left(\sqrt{m_2} + \sqrt{\operatorname{rank}(V)}\right)^2} \tag{2.11}$$

with some constants $C_1 < 1$ and $C_2 > 1$. In [GK17] we also use a condition similar to (2.11):

$$\operatorname{rank}(V A_0) \leq \frac{C_1(n\,m_2 - 1)}{C_2\left(\sqrt{m_2} + \sqrt{\operatorname{rank}(V)}\right)^2}. \tag{2.12}$$

Note that, as $\operatorname{rank}(V A_0) \leq \operatorname{rank}(A_0) \wedge \operatorname{rank}(V)$, condition (2.12) is weaker than (2.11).

In [GK17] we also provide empirical results that confirm our theoretical findings and illustrate the fact that using the Frobenius norm instead of the square Frobenius norm as a goodness-of-fit criterion makes the optimal smoothing parameter λ independent of the noise level, allowing for a better stability of the procedure with respect to the noise level, as compared to other state-of-the-art procedures.

2.5 One-bit matrix completion

In [KLMS15] jointly with Jean Lafond, Éric Moulines and Joseph Salmon we consider a statistical model where instead of observing a real-valued entry of an unknown matrix we are now able to see only highly quantized outputs. These discrete observations are generated according to a probability distribution which is parametrized by the corresponding entry of the unknown low-rank matrix.

The problem of matrix completion over a finite alphabet has received much less attention than the traditional unquantized matrix completion. This model is well suited for the analysis of voting patterns, preference ratings, or recovery of incomplete survey data, where typical survey responses are of the form "true/false", "yes/no" or "agree/disagree/no opinion" for instance.

One-bit matrix completion, corresponding to the case of binary, i.e. yes/no, observations, was introduced by [23] where the first theoretical guarantees on the performance of a nuclear-norm constrained maximum likelihood estimator are given. The sampling model considered in [23] assumes that the entries are sampled uniformly at random. Unfortunately, this condition is unrealistic for recommender system applications: in such a context some users are more active than others and popular items are rated more frequently. Another important issue is that the method of [23] requires the knowledge of an upper bound on the nuclear norm or on the rank of the unknown matrix. Such information is usually not available in applications.

One-bit matrix completion was further considered by [10] where a max-norm constrained maximum likelihood estimate is proposed. This method allows more

general non-uniform sampling schemes but still requires an upper bound on the max-norm of the unknown matrix. The rates of convergence obtained in [23] and [10] are slower than the rate of convergence of our estimator.

We consider the maximum likelihood estimator with nuclear-norm penalization. Our method allows us to consider general sampling schemes and only requires the knowledge of an upper bound on the maximum absolute value of the entries of the unknown matrix. All the previous works on this model also required the knowledge of this bound with sometimes the need of additional (and more difficult to obtain) information on the unknown matrix.

Assume that the observations follow a Bernoulli distribution parametrized by a matrix $\bar{X} \in \mathbb{R}^{m_1 \times m_2}$ and that an *i.i.d.* sequence of coefficients $(\omega_i)_{i=1}^n \in ([m_1] \times [m_2])^n$ is revealed. The observations associated to these coefficients are denoted by $(Y_i)_{i=1}^n \in \{1,2\}^n$ and distributed as follows

$$\mathbb{P}(Y_i = j) = f^j(\bar{X}_{\omega_i}), \quad j \in \{1,2\},$$

where $f = (f^j)_{j=1}^2$ is a $2-$link function. For example, taking $f^1(x) = \frac{e^x}{1+e^x}$ and $f^2(x) = 1 - f^1(x)$ we get the usual logistic regression. Here, the corresponding entries of $\bar{X}$ represent the log odds of the Bernoulli distribution that governs our observations. We have two goals: the first is the recovery of the distribution of Y given by $f(\bar{X})$ and, the second, is to accurately recover the matrix $\bar{X}$ itself.

In order to simplify notation, we write $\bar{X}_i$ instead of $\bar{X}_{\omega_i}$. Denote by Φ_{Y} the (normalized) negative log-likelihood of the observations:

$$\Phi_{\mathrm{Y}}(X) = -\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{2} \mathbb{1}_{\{Y_i = j\}} \log\left(f^j(\bar{X}_i)\right) \right).$$

Let $\gamma > 0$ be an upper bound of $\|\bar{X}\|_\infty$. In [KLMS15] we introduce the following estimator of $\bar{X}$:

$$\hat{X} = \underset{X \in \mathbb{R}^{m_1 \times m_2}, \|X\|_\infty \leq \gamma}{\arg\min} \Phi_Y^\lambda(X), \quad \text{where} \quad \Phi_Y^\lambda(X) = \Phi_{\mathrm{Y}}(X) + \lambda \|X\|_*,$$

with $\lambda > 0$ being a regularization parameter. We need the following assumptions which allows us to control the "steepness" and "flatness" of f:

Assumption 9. *We assume that the functions $x \mapsto -\log(f^j(x))$, $j = 1,2$ are convex. In addition, we suppose that there exist positive constants H_γ, L_γ and K_γ such that:*

$$H_\gamma \geq 2 \sup_{|x| \leq \gamma} (|\log(f^1(x))| \vee |\log(f^2(x))|), \tag{2.13}$$

$$L_\gamma \geq \max\left(\sup_{|x| \leq \gamma} \frac{|(f^1)'(x)|}{f^1(x)}, \sup_{|x| \leq \gamma} \frac{|(f^2)'(x)|}{f^2(x)} \right), \tag{2.14}$$

$$K_\gamma = \inf_{|x| \leq \gamma} g(x), \qquad \textit{where } g(x) = \frac{(f^1)'(x)^2}{8 f^1(x)(1 - f^1(x))}. \tag{2.15}$$

This assumption is quite mild. It includes, for example, the logistic regression (with $f^1(x) = \frac{e^x}{1+e^x}$) and the probit model (with $f^1(x) = \Phi(x/\sigma)$ where Φ is the cumulative distribution function of a standard Gaussian).

For this estimator we establish upper bounds both on the Frobenius norm between the unknown true matrix and the proposed estimator and on the associated Kullback-Leibler divergence. The former is addressed by Theorem 10 which provides an upper bound on the KL divergence between $f(\bar{X})$ and $f(\hat{X})$, the latter is tackled by Corollary 11, which bounds the Frobenius norm of $\bar{X}$ estimation error.

Theorem 10. *Let Assumptions 3, 4 and 9 be satisfied. Suppose that* $\|\bar{X}\|_\infty \leq \gamma$, $n \geq 2m\log(d)/(9\nu)$ *and take*

$$\lambda = 6L_\gamma \sqrt{\frac{2\nu \log(d)}{mn}}.$$

Then, with probability at least $1-3d^{-1}$ *the Kullback-Leibler divergence is bounded by*

$$\mathrm{KL}\left(f(\bar{X}), f(\hat{X})\right) \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \mathrm{rank}(\bar{X})}{K_\gamma} \frac{(m_1 \vee m_2)\log(d)}{n}, eH_\gamma \sqrt{\frac{\log(d)}{n}}\right),$$

with $\bar{c}$ *a universal constant whose value is specified in the proof.*

This result immediately gives an upper bound on the estimation error of $\hat{X}$, measured in Frobenius norm:

Corollary 11. *Under the assumptions of Theorem 10, we have that, with probability at least* $1 - 3d^{-1}$,

$$\frac{\|\bar{X} - \hat{X}\|_2^2}{m_1 m_2} \leq \mu \max\left(\bar{c}\mu\nu \frac{L_\gamma^2 \mathrm{rank}(\bar{X})}{K_\gamma^2} \frac{(m_1 \vee m_2)\log(d)}{n}, \frac{H_\gamma}{K_\gamma} \sqrt{\frac{\log(d)}{n}}\right).$$

We also establish lower bounds, showing that our upper bounds are minimax optimal up to logarithmic factors. Another contribution of [KLMS15] is an extension of one-bit matrix completion to the case of a more general finite alphabet. We also present an implementation based on the lifted coordinate descent algorithm introduced in [27] and Monte Carlo experiments supporting our claims.

2.6 Matrix completion by singular value thresholding: minimax optimal bounds.

Quite popular tool in the matrix completion literature are the thresholding methods which can be divided in two groups: one-step thresholding methods and iterative thresholding methods. Strong theoretical guarantees were obtained

for one-step thresholding procedures. For example, Koltchinskii et al in [38] introduce a soft-thresholding method and show that it is minimax optimal up to a logarithmic factor. In [Klo11] I consider a hard thresholding proceedure. In [20] Chatterjee proposes an universal singular value thresholding that can be applied to a large number of matrix estimation problems, including matrix completion. Despite strong theoretical guarantees, these one-step thresholding methods have two important drawbacks: they show poor behavior in practice and only work under the uniform sampling distribution which is not realistic in many practical situations.

Much better practical performances have been shown by iterative thresholding methods (see, e.g., [9, 44, 24]). For example, in [9], Cai et al propose a first-order singular value thresholding algorithm SVT which approximately solves the nuclear norm minimization problem. In [44], Mazmuder et al introduce `softImpute` algorithm. `SoftImpute` produces a sequence of solutions that converges to a solution of the nuclear norm regularized least-squares problem when the number of iterations goes to infinity. More recently Dhanjal et al [24] propose an improvement for the `softImpute` algorithm using randomized SVDs along with a novel updating method. This improvement allows to bypass the bottleneck in the algorithm which consists in the use of the singular value decomposition of a large matrix at each iteration.

These iterative thresholding algorithms are simple to implement, scale to relatively large matrices and achieve competitive errors compared to the state-of-the-art algorithms. On the other hand, the majority of existing algorithms for matrix completion consists of batch methods, that is, they operate on the full data matrix. However, in some applications, such as recommendation systems or localization in sensor networks, we observe a sequence of data matrix $M_1, \dots, M_T$ revealed sequentially where from M_t to M_{t+1} we add new observations. In such situations the predictive rule should be refined incrementally. One advantage of iterative thresholding algorithms is that they can be adapted to such sequential learning, see, for example, [24].

In spite of their empirical success, the theoretical guarantees of such iterative thresholding methods are poorly understood. In [Klo15] I provide strong theoretical guarantees, similar to those obtained for nuclear-norm penalization methods (see, for example, [46], [Klo14]) and one step thresholding methods (see [38, 20], [Klo11]) for a modification of the `softImpute` algorithm.

2.6.1 Algorithm

In [Klo15] I consider the *Bernoulli model* introduced in 2.1. We assume that the noise variables ε_{ij} are independent, zero mean and bounded:

Assumption 12. $\mathbb{E}(\varepsilon_{ij}) = 0$, $\mathbb{E}(\varepsilon_{ij}^2) = \sigma^2$ *and there exists a positive constant* $b > 0$ *such that*

$$\max_{i,j} |\varepsilon_{ij}| \leq b.$$

Our algorithm is based on the `softImpute` algorithm proposed by Mazumder et al [44] which is inspired by SVD-Impute of Troyanskaya et al [53]. It alternates

between imputing the missing values from a current SVD, and updating the SVD using the data matrix.

Algorithm 1

Require : Matrix Y, regularization parameter λ and a, an upper bound on the sup-norm of A_0.

1. $A^{old} = 0$
2. (a) Repeat
   (i) Compute $A^{new} \leftarrow S_\lambda\left(Y + (A^{old})_{\bar{\Omega}}\right)$.
   (ii) If $\left\|\left(A^{new} - A^{old}\right)_{\bar{\Omega}}\right\| < \lambda/3$ and $\left\|A^{new} - A^{old}\right\|_\infty < a$ exit.
   (iii) Put $A^{old} = \left(A_{ij}^{old}\right)$

$$A_{ij}^{old} = \begin{cases} A_{ij}^{new} & \text{if } |A_{ij}^{new}| \leq a \\ a & \text{if } A_{ij}^{new} > a \\ -a & \text{if } A_{ij}^{new} < -a. \end{cases}$$

   (b) Assign $\hat{A} \leftarrow A^{new}$.
3. Output $\hat{A}$.

This algorithm repeatedly replaces the missing entries with the current guess, update the guess by solving

$$A^{new} \in \underset{A}{\text{minimize}}\, f_\lambda(A) = \frac{1}{2}\|Y + (A^{old})_{\bar{\Omega}} - A\|_2^2 + \lambda\|A\|_*$$

and truncating A^{new}. In [Klo15] I show that this algorithm converges and I derive an upper bound on the estimation error of $\hat{A}$ produced by Algorithm 1 in the case of general sampling schemes. In order to compare this result with previous results on noisy matrix completion let me consider a more restrictive assumption on the sampling distribution. That is, I will assume that this distribution is close to the uniform one:

Assumption 13. *There exists positives constants μ_1 and μ_2 independent on m_1 and m_2 and a $0 < p < 1$ such that for every $(i, j) \in \{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$ we have*

$$\mu_2 p \leq \pi_{ij} \leq \mu_1 p.$$

Under this assumption the results of [Klo15] yield the following upper bound on the estimation risk of $\hat{A}$ produced by Algorithm 1 :

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \leq \frac{C(a \vee b)^2 \operatorname{rank}(A_0)}{pm}. \tag{2.16}$$

This bound holds with high probability and is non-asymptotic. In particular, it implies that the proposed estimator is minimax optimal in this setting.

Our model was previously considered by Chatterjee [20] in the case of uniform sampling distribution. The rate of convergence obtained in [20] is the following one:

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \leq C\sqrt{\frac{\operatorname{rank}(A_0)}{pm}}.$$

which is the square root of the optimal rate (2.16).

In [46, 38], [Klo14] a closely related set up using the trace regression model was considered. Putting $n = pm_1m_2$, we see that, compared to (2.16), bounds obtained in these papers contain an additional logarithmic factor $\log(m_1 + m_2)$. Koltchinskii et al in [38] obtained lower bounds for the estimation error without this additional $\log(m_1 + m_2)$ factor. So the results of [Klo15] answer an important theoretical question that of the exact minimax rate of convergence for the matrix completion problem. As the lower bound in [38] is obtained for a different setting, in [Klo15] I adapt their proof to the Bernoulli model and show that the minimax rate of convergence for matrix completion problem is given by (2.16) and that the estimator produced by our algorithm is minimax optimal.

2.7 Robust Matrix Completion

Matrix completion problem is motivated by a variety of applications. An important question in applications is whether or not matrix completion procedures are robust to corruptions. Suppose that we observe noisy entries of $A_0 = L_0 + S_0$ where L_0 is an unknown low-rank matrix and S_0 corresponds to some gross/malicious corruptions. We assume that S_0 has some low complexity structure such as entry-wise sparsity or column-wise sparsity. We wish to recover L_0, but only observe a few entries of A_0 and, among those, a fraction happens to be corrupted by S_0. Of course, we do not know which entries are corrupted.

It has been shown empirically that uncontrolled and potentially adversarial gross errors that might affect only a few observations are particularly harmful. For example, Xu et al [58] showed that a very popular matrix completion procedure using nuclear norm minimization can fail dramatically even if S_0 contains only a single nonzero column. It is particularly relevant in applications to recommendation systems where malicious users try to manipulate the prediction of matrix completion algorithms by introducing spurious perturbations S_0. Hence, the need for new techniques robust to the presence of corruptions S_0.

With these motivations in mind, in [KLT16] jointly with Karim Lounici and Alexandre Tsybakov we consider *robust matrix completion*. Assume that we observe $(X_i, Y_i), i = 1, \dots, N$ satisfying the trace regression model (2.1) and suppose that the set of observations is the union of two components Ω and $\tilde{\Omega}$ with $\Omega \cap \tilde{\Omega} = \emptyset$. The sets Ω and $\tilde{\Omega}$ are assumed non-random. One of them, Ω, corresponds to the "non-corrupted" observations of noisy entries of L_0, i.e. observations for which the corresponding entry of S_0 is zero. The other component, $\tilde{\Omega}$, corresponds to the observations for which the corresponding entry of S_0 is non vanishing. Given an observation, we do not know if it belongs to the corrupted or non-corrupted part of observations and we have that $|\Omega| + |\tilde{\Omega}| = N$.

A particular case of this setting is the matrix decomposition problem where $N = m_1 m_2$, i.e., we observe all entries of A_0. Several recent works consider the matrix decomposition problem, mostly in the noiseless setting, $\varepsilon_i \equiv 0$. Chandrasekaran et al. [19] analyzed the case when the matrix S_0 is sparse, with small number of non-zero entries. They proved that exact recovery of (L_0, S_0) is possible with high probability under additional identifiability conditions. This model was further studied by Hsu et al. [35] who give milder conditions for the exact recovery of (L_0, S_0). Also in the noiseless setting, Candes et al. [13] studied the same model but with positions of corruptions chosen uniformly at random. Xu et al. [58] studied a model, in which the matrix S_0 is columnwise sparse with sufficiently small number of non-zero columns. Their method guarantees approximate recovery for the non-corrupted columns of the low-rank component L_0. Agarwal et al. [2] consider a general model, in which the observations are noisy realizations of a linear transformation of A_0. Their setup includes the matrix decomposition problem and some other statistical models of interest but does not cover the matrix completion problem. component L_0. Their analysis includes as particular cases both the entrywise corruptions and the columnwise corruptions.

The robust matrix completion setting, when $N < m_1 m_2$, was first considered by Candes et al. [13] in the noiseless case for entrywise sparse S_0. They assumed that the support of S_0 is selected uniformly at random and that N is equal to $0.1 m_1 m_2$ or to some other fixed fraction of $m_1 m_2$. Chen et al. [21] considered also the noiseless case but with columnwise sparse S_0. They proved that the same procedure as in [19] can recover the non-corrupted columns of L_0 and identify the set of indices of the corrupted columns. More recently, Chen et al. [22] and Li [39] considered noiseless robust matrix completion with entrywise sparse S_0. They proved exact recovery of the low-rank component under an incoherence condition on L_0 and some additional assumptions on the number of corrupted observations.

To the best of my knowledge, [KLT16] is the first study of robust matrix completion with noise. Our analysis is general and covers in particular the cases of columnwise sparse corruptions and entrywise sparse corruptions. It is important to note that we do not require strong assumptions on the unknown matrices, such as the incoherence condition, or additional restrictions on the number of corrupted observations as in the noiseless case. This is due to the

fact that we do not aim at exact recovery of the unknown matrix. We emphasize that we do not need to know the rank of L_0 nor the sparsity level of S_0. We do not need to observe all entries of A_0 either.

Another important point is that our method allows us to consider quite general and unknown sampling distribution. All the previous works on noiseless robust matrix completion assume the uniform sampling distribution. However, in practice the observed entries are not guaranteed to follow the uniform scheme and the sampling distribution is not exactly known. We also prove the minimax lower bounds showing that the rates attained by our estimator are minimax optimal up to a logarithmic factor.

2.7.1 Column-wise sparsity

In the usual matrix completion ($S_0 = 0$) one of the most popular method is based on constrained nuclear norm minimization. We introduce an additional norm-based penalty that accounts for corruptions induced by the matrix S_0. This penalty depends on the structure of S_0. Suppose first that S_0 has at most s non-zero columns. Then, we use a $\|\cdot\|_{2,1}$ regularization:

$$(\hat{L}, \hat{S}) \in \underset{\substack{\|L\|_\infty \le a \\ \|S\|_\infty \le a}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \langle X_i, L + S \rangle\right)^2 + \lambda_1 \|L\|_1 + \lambda_2 \|S\|_{2,1} \right\} \tag{2.17}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are regularization parameters and a is an upper bound on $\|L_0\|_\infty$ and $\|S_0\|_\infty$. Our estimator requires an upper bound on the maximum of the absolute values of the entries of L_0 and S_0. Such information is often available in applications; for example, in recommendation systems, this bound is just the maximum rating.

In the case of column-wise corruptions, we show the following bound on the normalized Frobenius error of the estimator $(\hat{L}, \hat{S})$ (2.17) of (L_0, S_0): with high probability

$$\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \lesssim \frac{r\,(m_1 \vee m_2) + |\tilde{\Omega}|}{|\Omega|} + \frac{s}{m_2} \tag{2.18}$$

where the symbol $\lesssim$ means that the inequality holds up to a logarithmic factor and a multiplicative constant which may depend only on a and σ. Here, r denotes the rank of L_0, s is the number of corrupted columns, $|\Omega|$ and $|\tilde{\Omega}|$ are respectively the number of non-corrupted and corrupted observations.

In the upper bound (2.18) we have two terms. The first term, proportional to $r(m_1 \vee m_2)$, also appears in the usual matrix completion see, e.g., (2.16). The second one is induced by corruptions and is proportional to the number of corrupted columns s and to the number of corrupted observations, $|\tilde{\Omega}|$. This term is equal to zero if there is no corruption.

Suppose that the number of corrupted columns is small ($s \ll m_2$). Then, (2.18) guarantees, that the prediction error of our estimator is small whenever

the number of non-corrupted observations n satisfies the following condition

$$n \gtrsim (m_1 \vee m_2)\text{rank}(L_0) + |\tilde{\Omega}| \tag{2.19}$$

where $|\tilde{\Omega}|$ is the number of corrupted observations. This quantifies the sample size sufficient for successful, robust to corruptions, matrix completion. When $\text{rank}(L_0)$ is small and $s \ll m_2$, the right hand side of (2.19) is considerably smaller than the total number of entries $m_1 m_2$.

2.7.2 Element-wise sparsity

We consider now the case when S_0 has s non-zero entries but they do not necessarily lie in a small subset of columns. Here we use a l_1 regularization:

$$(\hat{L}, \hat{S}) \in \underset{\substack{\|L\|_\infty \le \mathbf{a} \\ \|S\|_\infty \le \mathbf{a}}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \langle X_i, L + S \rangle\right)^2 + \lambda_1 \|L\|_* + \lambda_2 \|S\|_1 \right\}. \tag{2.20}$$

In the case of element-wise corruptions, we show the following bound on the normalized Frobenius error of the estimator $(\hat{L}, \hat{S})$ (2.20) of (L_0, S_0): with high probability

$$\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \lesssim \frac{r\,(m_1 \vee m_2) + |\tilde{\Omega}|}{|\Omega|} + \frac{s}{m_1 m_2} \tag{2.21}$$

where the symbol $\lesssim$ means that the inequality holds up to a logarithmic factor and a multiplicative constant. As in the column sparsity case, we observe two terms in the upper bound (2.21). The first term, proportional to $r\,(m_1 \vee m_2)/n$, also appears in the usual matrix completion. The second and the third ones are induced by corruptions and are proportional to the number of nonzero entries in S_0, s, and to the number of corrupted observations, $|\tilde{\Omega}|$. When $s \ll n < m_1 m_2$, this bound implies that one can recover a low-rank matrix from a nearly minimal number of observations even when a part of these samples has been corrupted.

2.8 Estimation of matrices with row sparsity

In recent years, there has been a great interest for the theory of estimation in high-dimensional statistical models under different sparsity scenarii. The main motivation behind sparse estimation is based on the observation that, in several practical applications, the number of variables is much larger than the number of observations, but the degree of freedom of the underlying model is relatively small. One example of such sparse estimation is the problem of estimating of a sparse regression vector from a set of linear measurements. Another example is the problem of matrix recovery under the assumption that the unknown matrix has low rank.

In some recent papers dealing with covariance matrix estimation, a different notion of sparsity was considered (see, for example, [11], [50]). This notion is

based on sparsity assumptions on the rows (or columns) $A_{i\cdot}$ of a matrix A. One can consider the hard sparsity assumption meaning that each row $A_{i\cdot}$ of A contains at most s non-zero elements, or soft sparsity assumption, based on imposing a certain decay rate on ordered entries of $A_{i\cdot}$. These notions of sparsity can be defined in terms of l_q-balls for $q \in [0, 2)$:

$$\mathbb{B}_q(s) = \left\{ v = (v_i) \in \mathbb{R}^{m_2} \ : \ \sum_{i=1}^{m_2} |v_i|^q \le s \right\} \tag{2.22}$$

where $0 < s < \infty$ is a given constant. The case $q = 0$ corresponds to the set of vectors v with at most s non-zero elements:

$$\mathbb{B}_0(s) = \left\{ v = (v_i) \in \mathbb{R}^{m_2} \ : \ \sum_{i=1}^{m_2} \mathbb{I}(v_i \neq 0) \le s \right\} \tag{2.23}$$

here $\mathbb{I}(\cdot)$ denotes the indicator function and $s \ge 1$ is an integer.

In [KT15] jointly with Alexandre Tsybakov, we consider row sparsity setting in the matrix signal plus noise model. Suppose we have noisy observations $Y = (y_{ij})$ of a $m_1 \times m_2$ matrix $A = (a_{ij})$ where

$$y_{ij} = a_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, m_1, \quad j = 1, \dots, m_2,$$

here, ε_{ij} are i.i.d sub-Gaussian random variables. We study the minimax optimal rates of convergence for the estimation of A assuming that there exist $q \in [0, 2)$ and s such that $A_{i\cdot} \in \mathbb{B}_q(s)$ for any $i = 1, \dots, n_1$. The minimax rate of convergence characterizes the fundamental limitation of the estimation accuracy. It also captures the interdependence between the different parameters in the model. There is a rich line of work on such fundamental limits (see, for example, [36, 54]).

The minimax risk depends crucially on the choice of the norm in the loss function. In [KT15] we measure the estimation error in $\|\cdot\|_{2,p}$-(quasi)norm: for any $A = (A_{1\cdot}, \dots, A_{n_1\cdot})^T \in \mathbb{R}^{n_1 \times n_2}$ and $p > 0$ we define

$$\|A\|_{2,p} = \left(\sum_{i=1}^{n_1} \|A_{i\cdot}\|_2^p \right)^{1/p}. \tag{2.24}$$

Note that when $m_1 = 1$, we obtain the problem of estimating of a vector $v = (v_i) \in \mathbb{B}_q(s) \subset \mathbb{R}^{m_2}$ from noisy observations:

$$y_i = v_i + \varepsilon_i, \quad i = 1, \dots, m_2.$$

Here ε_{ij} are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. This problem was considered in a number of papers (see, for example, [26, 6, 1, 49]). Let η_{vect} denotes the minimax rate of convergence in the vector case. For instance, the non-asymptotic minimax optimal rate of convergence for estimation of v in the l_2-norm, obtained in [6], is given by

$$\eta_{vect}(s) = \sigma^2 \, s \, \log\left(\frac{e \, m_2}{s}\right)$$

when $q = 0$ and by

$$\eta_{vect}(s) = \left(s \left[\sigma^2 \log \left(1 + \frac{\sigma^q m_2}{s} \right) \right]^{1-q/2} \right) \vee \left(s^{2/q} \right) \vee \left(m_2 \sigma^2 \right)$$

when $0 < q < 2$. Results obtained in [KT15] imply that the minimax rate of convergence for the estimation of matrices under the row sparsity assumption is $m_1 \times \eta_{vect}$. Thus, the problem reduces to estimating each row separately. The additional matrix structure does not lead to improvement or deterioration of the rate of convergence. A major focus in [KT15] is on the derivation of lower bounds, which is a key step in establishing minimax optimal rates of convergence.

Chapter 3

Varying Coefficient Model

One of the fundamental tasks in statistics is to characterize the relationship between a set of covariates and a response variable. Varying coefficient model is commonly used for describing time-varying covariate effects and represents a useful tool for exploring dynamic patterns in economics, epidemiology, ecology, etc. It provides a more flexible approach than the classical linear regression model and is often used to analyse the data measured repeatedly over time.

Let (W_i, t_i, Y_i), $i = 1, \dots, n$ be sampled independently from the varying coefficient model

$$Y = W^T f(t) + \sigma\xi. \tag{3.1}$$

Here, $W \in \mathbb{R}^p$ are random vectors of predictors, $f(\cdot) = (f_1(\cdot), \dots, f_p(\cdot))^T$ is an unknown vector-valued function of regression coefficients and $t \in [0, 1]$ is a random variable independent of W. Let μ denote its distribution. The noise variable ξ is sub-exponential, independent of W and t and such that $\mathbb{E}(\xi) = 0$ and $\mathbb{E}(\xi^2) = 1$, $\sigma > 0$ denotes the noise level. The goal is to estimate the vector function $f(\cdot)$ on the basis of observations (W_i, t_i, Y_i), $i = 1, \dots, n$.

The varying coefficient models were introduced by Cleveland, Grosse and Shyu [55] and Hastie and Tibshirani [34] and have been extensively studied in the past 15 years. Existing methods typically provide asymptotic evaluation of the precision of the estimation procedures under the assumption that the number of observations tends to infinity. In practical applications, however, only a finite number of measurements are available. In [KP13] and [KP15] jointly with Marianna Pensky we focus on a non-asymptotic approach to this problem.

3.1 Nuclear norm minimization approach

The estimation method proposed in [KP13] is based on the approximation of the unknown functions $f_i(t)$ using a basis expansion. This approximation generates the coordinate matrix A_0. In the above model, some of the components of the vector function f are constant. The larger the part of the constant regression

coefficients, the smaller the rank of the coordinate matrix A_0 (the rank of the matrix A_0 does not exceed the number of time-varying components of vector $f(\cdot)$ by more than one). We suppose that the first element of this basis is just a constant function on $[0,1]$. In this case, if the component $f_i(\cdot)$ is constant, then, it has only one non-zero coefficient in its expansion over the basis. This suggests the idea to take into account the number of constant regression coefficients using the rank of the coordinate matrix A_0.

In [KP13] we propose a novel estimation procedure which is based on recent developments in matrix estimation. Our procedure involves estimating A_0 using nuclear-norm penalization which is now a well-established proxy for rank penalization in the compressed sensing literature.

The first step of our estimation method is the approximation of the unknown functions $f_i(t)$ by expanding them over an appropriate basis. The possible choices of the basis include the standard Fourier basis and wavelets. Let $(\phi_i(\cdot))_{i=1,\dots,\infty}$ be an orthonormal basis in $L_2((0,1),d\mu)$, $l \in \mathbb{N}$ and $\phi(\cdot) = (\phi_1(\cdot),\dots,\phi_l(\cdot))^T$. We assume that basis functions satisfy the following condition: there exists $c_\phi < \infty$ such that

$$\left\|\phi^T(t)\right\|_2^2 = \sum_{j=1}^{l} |\phi_j(t)|^2 \le c_\phi^2\, l, \tag{3.2}$$

for any $l \ge 1$ and any $t \in [0,1]$. Note that this condition is satisfied for most of the usual bases. We introduce the coordinate matrix $A_0 \in \mathbb{R}^{p\times l}$ with elements

$$a_{kj}^0 = \langle f_k, \phi_j\rangle_{L_2(d\mu)}, \quad k = 1,\cdots,p,\ j = 1,\cdots,l.$$

For each $k = 1,\dots,p$, we have

$$f_k(t) = \sum_{j=1}^{l} a_{kj}^0 \phi_j(t) + \rho_k^{(l)}(t). \tag{3.3}$$

Denote the remainder by $\rho^{(l)}(\cdot) = (\rho_1^{(l)}(\cdot),\dots,\rho_p^{(l)}(\cdot))^T$. We assume that f_k is well approximated by $\sum_{j=1}^{l} a_{kj}^0\phi_j(t)$:

Assumption 14. *We assume that the basis satisfies condition (3.2) and that there exists a positive constant b such that, for any $l \ge 1$*

$$\left\|\rho^{(l)}(\cdot)\right\|_\infty \quad \le \quad b\, l^{-\gamma}, \quad \gamma > 0. \tag{3.4}$$

Often approximation in L_2-norm gives better rates of convergence. In order to get upper bounds on the mean squared error we will use the following additional assumption:

Assumption 15. *There exists $b_1 > 0$ such that, for any $l \ge 1$*

$$\left\|\rho^{(l)}(\cdot)\right\|_{L_2(d\mu)} \le b_1\, l^{-(\gamma+1/2)}, \quad \gamma > 0.$$

We assume that the vectors W_i are i.i.d copies of a random vector W having distribution Π on a given set of vectors $\mathcal{X}$. Using rescaling, we can suppose that $\|W\|_2 \leq 1$. Let $\mathbb{E}\left(W\,W^T\right) = \Omega$ and $\omega_{\max}$, $\omega_{\min}$ denote respectively its maximal and minimal singular values. We need the following assumption on the distribution of W:

Assumption 16. *The matrix $\Omega = \mathbb{E}\left(W\,W^T\right)$ is positive definite.*

We introduce the following notations:

$$\omega = \mathrm{tr}(\Omega) \vee (l\,\omega_{\max}) \quad \text{and} \quad n^{**} = \frac{C\,c_\phi^2\,l\,\log(d)}{\omega_{\min}^2}\left[(\omega\,s) \vee 1\right]..$$

Denoting $X = W\phi^T(t)$, we can rewrite (3.1) in the following form

$$Y = \mathrm{tr}\left(A_0 X^T\right) + W^T\rho^{(l)}(t) + \sigma\xi. \tag{3.5}$$

We suppose that some of the functions $f_i(\cdot)$ are constant and let $s-1$ denote the number of non-constant $f_i(\cdot)$. Note that $\mathrm{rank}\,(A_0) \leq s$. Based on the observations (Y_i, X_i), we define the following estimator of A_0:

$$\hat{A} = \underset{A\in\mathbb{R}^{p\times l}}{\arg\min}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \langle X_i, A\rangle\right)^2 + \lambda\,\|A\|_*\right\}, \tag{3.6}$$

where λ is the regularization parameter. Subsequently, the estimator of the coordinate matrix is plugged into the expansion yielding the estimator $\hat{f}(\cdot) = \left(\hat{f}_1(\cdot), \ldots, \hat{f}_p(\cdot)\right)^T$ of the vector function $f(t)$. For this estimator we obtain upper bounds on the mean squared error $\frac{1}{p}\sum\limits_{i=1}^{p}\|\hat{f}_i - f_i\|^2_{L_2(d\mu)}$ and on the pointwise estimation error $\frac{1}{p}\sum\limits_{i=1}^{p}|\hat{f}_i(t) - f_i(t)|$ for any $t \in \mathrm{supp}(\mu)$:

Theorem 17. *Let Assumptions 14, 16 and 1 hold. With probability greater than $1 - 4/d$, one has*

(a) $\forall t \in \mathrm{supp}(\mu)$

$$\frac{1}{p}\sum_{i=1}^{p}|\hat{f}_i(t) - f_i(t)| \leq \frac{C\,\|\phi(t)\|_2^2\,\beta}{n} + \frac{2\,b^2\,s}{p\,l^{2\gamma}},$$

(b) If, in addition, Assumption 15 holds

$$\frac{1}{p}\sum_{i=1}^{p}\|\hat{f}_i - f_i\|^2_{L_2(d\mu)} \leq \frac{C\,\beta}{n} + \frac{2\,b_1^2\,s}{p\,l^{(2\gamma+1)}},$$

where

$$\beta = \begin{cases} \left(\sigma^2 + \dfrac{b^2\,(s-1)}{l^{2\gamma}}\right)\dfrac{\omega\,s\,\log(d)}{p\,\omega_{\min}^2}, & \text{if} \quad n \geq n^{**} \\ \max\left\{\left(\sigma^2 + \dfrac{b^2\,(s-1)}{l^{2\gamma}} + l\,\|A_0\|_*^2\right)\dfrac{\omega\,s\,\log(d)}{p\,\omega_{\min}^2}, \dfrac{c_\phi\,\|A_0\|_*^2\,\sqrt{\log(d)\,l\,n}}{\omega_{\min}\,p}\right\}, & \text{if not.} \end{cases}$$

These oracle inequalities are non-asymptotic and hold for finite values of p and n.

3.2 Sparse high-dimensional varying coefficient model

More recently a few authors considered still asymptotic but high-dimensional approach to the problem. Here, we refer to the situation where both the number of unknown parameters and the number of observations are large and the former may be of much higher dimension than latter. For example, Wei *et al.* [56] applied group Lasso for variable selection, while Lian [40] used the extended Bayesian information criterion. Fan *et al.* [28] applied nonparametric independence screening. Their results were extended by Lian and Ma [41] to include rank selection in addition to variable selection.

One important aspect that has not been studied in the existing literature is the non-asymptotic approach to the estimation, prediction and variables selection in the high-dimensional varying coefficient model. Some interesting questions arise in this non-asymptotic setting. One of them is the fundamental question of the minimax optimal rates of convergence. Our joint paper with M. Pensky [KP15] presents the first non-asymptotic minimax study of the sparse heterogeneous varying coefficient model.

In [KP15], we consider the case when the solution is sparse, in particular, only a few of the covariates are present and only some of them are time dependent. This setup is close to the one studied in a recent paper of Liang [40]. One important difference, however, is that in [40], the estimator is not adaptive to the smoothness of the time dependent covariates. In addition, Liang [40] assumes that all time dependent covariates have the same degree of smoothness and are spatially homogeneous. On the contrary, we consider a much more flexible and realistic scenario where the time dependent covariates possibly have different degrees of smoothness and may be spatially inhomogeneous.

Modern technologies produce very high dimensional data sets and, hence, stimulate an enormous interest in variable selection and estimation under a sparse scenario. In such scenarios, penalization-based methods are particularly attractive. Significant progress has been made in understanding the statistical properties of these methods. For example, many authors have studied the variable selection, estimation and prediction properties of the LASSO in the high-dimensional setting. A related LASSO-type procedure is the group-LASSO, where the covariates are assumed to be clustered in groups.

In order to construct a minimax optimal estimator, we introduce the block Lasso which can be viewed as a version of the group LASSO. However, unlike in group LASSO, where the groups occur naturally, the blocks in block LASSO are driven by the need to reduce the variance as it is done, for example, in block thresholding. Note that our estimator does not require the knowledge of which of the covariates are indeed present and which are time dependent. It adapts to

sparsity, to heterogeneity of the time dependent covariates and to their possibly spatial inhomogeneous nature.

We start by using the basis expansion described in Section 3.1. Then, for each function f_j, $j = 1, \cdots, p$, we divide its coefficients into $M + 1$ different groups where group zero contains only coefficient a_{j0} for the constant function $\phi_0(t) = 1$ and M groups of size $d \approx \log n$ where $M = l/d$. We denote $\mathbf{a}_{j0} = a_{j0}$ and $\mathbf{a}_{ji} = (a_{j,d(i-1)+1}, \cdots, a_{j,di})^T$ the sub-vector of coefficients of the function f_j in block i, $i = 1, \cdots, M$. We define the block norm of the matrix $\mathbf{A}$ as follows

$$\|\mathbf{A}\|_{\text{block}} = \sum_{j=1}^{p} \sum_{i=0}^{M} \|\mathbf{a}_{ji}\|_2. \tag{3.7}$$

Observe that $\|\mathbf{A}\|_{\text{block}}$ indeed satisfies the definition of a norm and is a sum of absolute values of coefficients a_{j0} of functions f_j and l_2 norms for each of the block vectors of coefficients $\mathbf{a}_{ji}$, $j = 1, \cdots, p$, $i = 1, \cdots, M$.

The penalty which we impose is related to both the ordinary and the group LASSO penalties which have been used by many authors. The difference, however, lies in the fact that the structure of the blocks is not motivated by naturally occurring groups (like, e.g., rows of the matrix $\mathbf{A}$) but rather our desire to exploit sparsity of functional coefficients a_{ji}.

In [KP15], we construct an estimator $\widehat{\mathbf{A}}$ of A_0 as a solution of the following convex optimization problem

$$\widehat{\mathbf{A}} = \arg\min_{\mathbf{A}} \left\{ n^{-1} \sum_{i=1}^{n} \left[Y_i - \text{Tr}(\mathbf{A}^T \phi(t_i) W_i^T) \right]^2 + \delta \|\mathbf{A}\|_{block} \right\}, \tag{3.8}$$

where δ is the regularisation parameter. Subsequently, we construct an estimator $\hat{\mathbf{f}}(t) = (\hat{f}_1(t), \cdots, \hat{f}_p(t))^T$ of the vector function $\mathbf{f}(t)$ using

$$\hat{f}_j(t) = \sum_{k=0}^{l} \hat{a}_{jk} \phi_k(t), \quad j = 1, \cdots, p. \tag{3.9}$$

In [KP15], we derive an upper bound for the risk of the estimator $\widehat{\mathbf{A}}$. In order to obtain a benchmark of how well the procedure is performing, we also determine lower bounds for the risk of any estimator $\widehat{\mathbf{A}}$ and show that our estimator attains those bounds within a constant (if all time-dependent covariates are spatially homogeneous) or a logarithmic factor of the number of observations.

Chapter 4

Network models and sparse graphon estimation

A network model is a natural way of representing objects and their relationships. Biological, social, technological networks can be studied as graphs, and the analysis of random graphs has become crucial in order to understand features of these systems.

The number of vertices n and the number of edges $|E|$ are main graph integral characteristics. The relation between these two parameters determines whether graph is sparse or dense. The maximal number of edges in a simple undirected graph is proportional to n^2. A dense graph is a graph in which the number of edges is close to this maximal number. Most real life networks are sparse. For instance, in social networks such as Facebook, each user is connected with a small group of friends rather then with the whole Facebook community. For sparse networks the number of edges is much smaller then the maximal possible number.

In general, sparse graphs are more difficult to handle than dense ones and difficulties increase as the graph get sparser. Even the almost dense case with $|E| \sim n^{2-o(1)}$ is rather different from the dense case. The extremely sparse case of graphs with bounded degree i.e. where all degrees are smaller than a fixed positive integer is very different having many novel features. Networks that occur in applications are usually between these two extremes of dense and bounded degree graphs. They often correspond to inhomogeneous networks with density of edges tending to 0 but with the maximum degree tending to infinity as n grows.

4.1 Network sequence model

In [KTV16] jointly with Alexandre Tsybakov and Nicolas Verzelen we consider a network defined as an undirected simple graph with n nodes. Assume that we observe the values $\mathbf{A}_{ij} \in \{0, 1\}$ where $\mathbf{A}_{ij} = 1$ is interpreted as the fact

that the nodes i and j are connected and $\mathbf{A}_{ij} = 0$ otherwise. We set $\mathbf{A}_{ii} = 0$ for all $1 \leq i \leq n$ and we assume that $\mathbf{A}_{ij}$ is a Bernoulli random variable with parameter $(\mathbf{\Theta}_0)_{ij} = \mathbb{P}(\mathbf{A}_{ij} = 1)$ for $1 \leq j < i \leq n$. The random variables $\mathbf{A}_{ij}$, $1 \leq j < i \leq n$, are assumed independent. We denote by $\mathbf{A}$ the adjacency matrix i.e., the $n \times n$ symmetric matrix with entries $\mathbf{A}_{ij}$ for $1 \leq j < i \leq n$ and zero diagonal entries. Similarly, we denote by $\mathbf{\Theta}_0$ the $n \times n$ symmetric matrix with entries $(\mathbf{\Theta}_0)_{ij}$ for $1 \leq j < i \leq n$ and zero diagonal entries. This is a matrix of probabilities associated to the graph; the nodes i and j are connected with probability $(\mathbf{\Theta}_0)_{ij}$. The model with such observations $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \leq j < i \leq n)$ is a special case of the inhomogeneous random graph model that, for definiteness, we will call the *network sequence model*[1], to emphasize the parallel with the Gaussian sequence model.

Our goal is to estimate the matrix of connexion probabilities $\mathbf{\Theta}_0$ under the Frobenius loss. In [KTV16] we are specially interested in the fundamental limits of estimation accuracy and our aim is to get minimax rates of convergence on suitable classes of matrices $\mathbf{\Theta}_0$. We have two cases: the case of dense graph when the maximum absolute value of the entries of $\mathbf{\Theta}_0$ is a fixed constant and the case of the sparse graph when connexion probabilities depend on n and go to zero when n goes to infinity.

The estimation of $\mathbf{\Theta}_0$ has already been considered by [20, 59, 18] but the convergence rates obtained there are far from being optimal. More recently, Gao et al. [31] have established the minimax estimation rates for $\mathbf{\Theta}_0$ on classes of block constant matrices and on the smooth graphon classes. Their analysis is restricted to the dense case. In [KTV16] we provide an extension for the sparse graphon model of minimax results in [31].

For example, consider the stochastic block models. Given an integer k and any $\rho_n \in (0, 1]$, let $\mathcal{T}[k, \rho_n]$ be the set of all probability matrices corresponding to k-class stochastic block model with connection probability uniformly smaller than ρ_n. Gao et al. [31] have proved that the minimax estimation rate over $\mathcal{T}[k, 1]$ is of the order $\frac{k^2}{n^2} + \frac{\log(k)}{n}$. The following theorem extends their results to an arbitrarily small $\rho_n > 0$:

Theorem 18. *Consider the network sequence model. For all $k \leq n$ and all $0 < \rho_n \leq 1$,*

$$\inf_{\widehat{\boldsymbol{T}}} \sup_{\mathbf{\Theta}_0 \in \mathcal{T}[k,\rho_n]} \mathbb{E}_{\mathbf{\Theta}_0} \left[\frac{1}{n^2} \|\widehat{\boldsymbol{T}} - \mathbf{\Theta}_0\|_F^2 \right] \approx \min \left(\rho_n \Big(\frac{\log(k)}{n} + \frac{k^2}{n^2} \Big), \rho_n^2 \right) \tag{4.1}$$

where $\mathbb{E}_{\mathbf{\Theta}_0}$ denotes the expectation with respect to the distribution of $\mathbf{A}$ when the underlying probability matrix is $\mathbf{\Theta}_0$ and $\inf_{\widehat{\boldsymbol{T}}}$ is the infimum over all estimators.

If $\rho_n \geq \frac{\log(k)}{n} + \frac{k^2}{n^2}$, the minimax rate of estimation is of the order $\rho_n \left(\frac{\log(k)}{n} + \frac{k^2}{n^2} \right)$. In [KTV16] we show that this rate is achieved by the restricted least squares estimator with $r \asymp \rho_n$ and by the least squares estimator if the partition

[1] In some recent papers, it is also called the inhomogeneous Erdös-Rényi model, which is somewhat ambiguous since the words "Erdös-Rényi model" designate a homogeneous graph.

is balanced and $\rho_n \geq k\log(k)/n$. For really sparse graphs, that is for ρ_n smaller than $\frac{\log(k)}{n} + \frac{k^2}{n^2}$, the estimation problem becomes rather trivial. Let $\overline{A} = 2\sum_{j<i} \mathbf{A}_{i,j}/(n(n-1))$ denote the edge density of the graph. Then, we have that, both the null estimator $\widehat{\boldsymbol{T}} = 0$ and the constant least squares estimator $\widehat{\boldsymbol{\Theta}}$ with all entries $\widehat{\boldsymbol{\Theta}}_{ij} = \overline{A}$ achieve the optimal rate ρ_n^2.

4.2 Graphon model

Actually one can consider a more general model which is called the graphon model. The reason for considering this more general model is the following one: real-life networks are in permanent movement and often their size is growing. Therefore, it is natural to look for a well-defined "limiting object" independent of the network size n and such that a stochastic network can be viewed as a partial observation of this limiting object. Such objects called *graphons* play a central role in the recent theory of graph limits introduced by Lovász and Szegedy [42]. For a detailed description of this beautiful theory we refer to the monograph by Lovász [43].

Graphons are symmetric measurable functions $W : [0,1]^2 \to [0,1]$ and every graph limit can be represented by a graphon. Graphons give a natural way of generating random graphs [43, 25]: the probability that two distinct nodes i and j are connected in the graphon model is the random variable

$$(\boldsymbol{\Theta}_0)_{ij} = W_0(\xi_i, \xi_j) \tag{4.2}$$

where $\xi_1, \ldots, \xi_n$ are unobserved (latent) i.i.d. random variables uniformly distributed on $[0,1]$. As above, the diagonal entries of $\boldsymbol{\Theta}_0$ are zero. Conditionally on $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$, the observations $\mathbf{A}_{ij}$ for $1 \leq j < i \leq n$ are assumed to be independent Bernoulli random variables with success probabilities $(\boldsymbol{\Theta}_0)_{ij}$.

For any positive integer n, a graphon function W_0 defines a probability distribution on graphs of size n. Note that this model is different from the network sequence model since the observations $\mathbf{A}_{ij}$ are no longer independent. If W_0 is a step-function with k steps, we obtain an analog of the stochastic block model with k groups. More generally, many exchangeable distributions on networks [25], including random dot product graphs [52] and some geometric random graphs [47] can be expressed as graphons.

It is easy to see that the expected number of edges in model (4.2) is a constant times the squared number of vertices, which corresponds to the dense case. For a given $\rho_n > 0$, one can modify the definition (4.2) to get a random graph model with $O(\rho_n n^2)$ edges. It is usually assumed that $\rho_n \to 0$ as $n \to \infty$. The adjacency matrix $\mathbf{A}'$ is sampled according to graphon $W_0 \in \mathcal{W}$ with scaling parameter ρ_n if for all $j < i$,

$$(\boldsymbol{\Theta}_0)_{ij} = \rho_n W_0(\xi_i, \xi_j). \tag{4.3}$$

The parameter ρ_n can be interpreted as the expected proportion of non-zero edges. Alternatively, model (4.3) can be considered as a graphon model (4.2)

that has been sparsified in the sense that its edges have been independently removed with probability $1-\rho_n$ and kept with probability ρ_n. This sparse graphon model was considered in [4, 5, 57, 59].

Given an observed adjacency matrix $\mathbf{A}'$ sampled according to the model (4.2), the graphon function W_0 is not identifiable. This is because the topology of a network is invariant with respect to any change of labelling of its nodes. Consequently, for any given function $W_0(\cdot,\cdot)$ and a measure-preserving bijection $\tau : [0,1] \to [0,1]$ (with respect to the Lebesgue measure), the functions $W_0(x,y)$ and $W_0^{\tau}(x,y) := W_0(\tau(x),\tau(y))$ define the same probability distribution on random graphs. This motivates considering the equivalence classes of graphons that are weakly isomorphic. The corresponding quotient space is denoted by $\widetilde{\mathcal{W}}$. For any estimator $\check{f}$ of $f_0 = \rho_n W_0$, we define the squared error in the following way:

$$\delta^2(\check{f}, f_0) := \inf_{\tau\in\mathcal{M}} \int\int_{(0,1)^2} \left|f_0(\tau(x),\tau(y)) - \check{f}(x,y)\right|^2 \mathrm{d}x\mathrm{d}y$$

where $\mathcal{M}$ is the set of all measure-preserving bijections $\tau : [0,1] \to [0,1]$. It has been proved in [43, Ch.8,13] that $\delta(\cdot,\cdot)$ defines a metric on the quotient space $\widetilde{\mathcal{W}}$ of graphons.

In order to contrast the problem of graphon estimation with the estimation of $\mathbf{\Theta}_0$, one can invoke an analogy with the random design nonparametric regression. Suppose that we observe (y_i,ξ_i), $i = 1,\ldots,n$, that are independently sampled according to the model $y = f(\xi) + \epsilon$ where f is an unknown regression function, ϵ is a zero mean random variable and ξ is distributed with some density h on $[0,1]$. Given a sample of (y_i,ξ_i), the estimation of f with respect to the *empirical* loss is equivalent to the estimation of the vector $(f(\xi_1),\ldots,f(\xi_n))$ in, for instance, the Euclidean norm. On the other hand, estimation under the integrated loss consists in constructing an estimator $\hat{f}$ such that the integral $\int(\hat{f}(t) - f(t))^2 h(t)dt$ is small. Following this analogy, estimation of $\mathbf{\Theta}_0$ corresponds to an empirical loss problem whereas the graphon estimation corresponds to an integrated loss problem. However, as opposed to nonparametric regression, in the graphon models (4.2) and (4.3) the design $\xi_1,\ldots,\xi_n$ is not observed, which makes it quite challenging to derive the convergence rates in these settings.

4.2.1 From probability matrix estimation to graphon estimation

To any $n \times n$ probability matrix $\mathbf{\Theta}$ we can associate a graphon. This provides a way of deriving an estimator of $f_0(\cdot,\cdot) = \rho_n W_0(\cdot,\cdot)$ from an estimator of $\mathbf{\Theta}_0$. Given a $n \times n$ matrix $\mathbf{\Theta}$ with entries in $[0,1]$, define the *empirical graphon* $\widetilde{f}_{\mathbf{\Theta}}$ as the following piecewise constant function:

$$\widetilde{f}_{\mathbf{\Theta}}(x,y) = \mathbf{\Theta}_{\lceil nx\rceil,\lceil ny\rceil} \tag{4.4}$$

for all x and y in $(0,1]$. For any $W_0 \in \mathcal{W}$, $\rho_n > 0$, and any estimator $\widehat{\boldsymbol{T}}$ of $\boldsymbol{\Theta}_0$ using the triangle inequality we get that

$$\mathbb{E}\left[\delta^2(\widetilde{f}_{\widehat{\boldsymbol{T}}}, f_0)\right] \le 2\,\mathbb{E}\left[\frac{1}{n^2}\|\widehat{\boldsymbol{T}} - \boldsymbol{\Theta}_0\|_F^2\right] + 2\,\mathbb{E}\left[\delta^2\left(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0\right)\right]. \tag{4.5}$$

The bound on the integrated risk in (4.5) is the sum of two terms. The first term containing $\|\widehat{\boldsymbol{T}} - \boldsymbol{\Theta}_0\|_F^2$ is the *estimation error* term. The second term containing $\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)$ measures the distance between the true graphon f_0 and its discretized version sampled at the unobserved random design points $\xi_1, \dots, \xi_n$. We call it the *agnostic error*. The behavior of $\delta^2(\widetilde{f}_{\boldsymbol{\Theta}_0}, f_0)$ depends on the topology of the considered graphons.

In [KTV16] we obtain δ-norm non-asymptotic rates for graphon estimation problem on classes of step functions (analogs of stochastic block models) and on classes of smooth graphons in model (4.3). For instance, define $\mathcal{W}[k]$ the collection of k-step graphons, that is the subset of graphons $W \in \mathcal{W}$ such that for some $\mathbf{Q} \in \mathbb{R}^{k\times k}_{\mathrm{sym}}$ and some $\phi : [0,1] \to [k]$,

$$W(x,y) = \mathbf{Q}_{\phi(x),\phi(y)} \qquad \text{for all } x, y \in [0,1]\ . \tag{4.6}$$

For the step function graphons, we get the following upper bound on the minimax risk:

$$\inf_{\widehat{f}} \sup_{W_0 \in \mathcal{W}[k]} \mathbb{E}_{W_0}\left[\delta^2\left(\widehat{f}, f_0\right)\right] \le C\left\{\left[\rho_n\left(\frac{k^2}{n^2} + \frac{\log(k)}{n}\right) + \rho_n^2\sqrt{\frac{k}{n}}\right] \wedge \rho_n^2\right\} \tag{4.7}$$

where C is an absolute constant. Here, $\mathbb{E}_{W_0}$ denotes the expectation with respect to the distribution of observations $\mathbf{A}' = (\mathbf{A}_{ij}, 1 \le j < i \le n)$ when the underlying sparse graphon is $\rho_n W_0$ and $\inf_{\widehat{f}}$ is the infimum over all estimators. In [KTV16] we show that the upper bound (4.7) is optimal in a minimax sense (up to a logarithmic factor in k in one of the regimes). The bounds (4.7) imply that there are three regimes depending on the sparsity parameter ρ_n:

(i) *Weakly sparse graphs:* $\rho_n \ge \frac{\log(k)}{\sqrt{kn}} \vee (\frac{k}{n})^{3/2}$. The minimax risk is of the order $\rho_n^2\sqrt{k/n}$, and thus it is driven by the agnostic error arising from the lack of knowledge of the design.

(ii) *Moderately sparse graphs:* $\frac{\log(k)}{n} \vee \left(\frac{k}{n}\right)^2 \le \rho_n \le \frac{\log(k)}{\sqrt{kn}} \vee \left(\frac{k}{n}\right)^{3/2}$. The risk bound (4.7) is driven by the probability matrix estimation error. The upper bound (4.7) is of the order $\rho_n\left(\frac{k^2}{n^2} + \frac{\log(k)}{n}\right)$, which is the optimal rate of probability matrix estimation, cf. Theorem 18. It is optimal up to $\log(k)$ factor with respect to the $\delta(\cdot,\cdot)$ distance.

(iii) *Highly sparse graphs:* $\rho_n \le \frac{\log(k)}{n} \vee \left(\frac{k}{n}\right)^2$. The minimax risk is of the order ρ_n^2, and it is attained by the null estimator.

In a work parallel to [KTV16], Borgs et al. [7] provide an upper bound for the risk of step function graphon estimators in the context of privacy. When there are no privacy issues, comparing the upper bound of [7] with the optimal rate (4.7), we see that it has a suboptimal rate, which is the square root of the rate given by (4.7) in the moderately sparse zone. Note also that the setting in [7] is restricted to balanced partitions while in [KTV16] we consider more general partitions.

List of Publications

[dRT03] R. del Rio and Olga Tchebotareva. Boundary conditions of Sturm-Liouville operators with mixed spectra. *J. Math. Anal. Appl.*, 288(2):518–529, 2003.

[dRT07] Rafael del Rio and Olga Tchebotareva. Sturm-Liouville operators in the half axis with local perturbations. *J. Math. Anal. Appl.*, 329(1):557–566, 2007.

[GK17] Stéphane Gaïffas and Olga Klopp. High dimensional matrix estimation with unknown variance of the noise. *to appear in Statistica Sinica*, 2017.

[KLMS15] O. Klopp, J. Lafond, E. Moulines, and J. Salmon. Adaptive multinomial matrix completion. *Electron. J. Statist.*, 9(2):2950–2975, 2015.

[Klo11] O. Klopp. Rank penalized estimators for high-dimensional matrices. *Electron. J. Statist.*, 5:1161–1183, 2011.

[Klo14] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014.

[Klo15] Olga Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electron. J. Stat.*, 9(2):2348–2369, 2015.

[KLT16] O. Klopp, K. Lounici, and A. B. Tsybakov. Robust Matrix Completion. *Probability Theory and Related Fields*, pages 1–42, 2016.

[KP13] Olga Klopp and Marianna Pensky. Non-asymptotic approach to varying coefficient model. *Electron. J. Stat.*, 7:454–479, 2013.

[KP15] Olga Klopp and Marianna Pensky. Sparse high-dimensional varying coefficient model: nonasymptotic minimax study. *Ann. Statist.*, 43(3):1273–1299, 2015.

[KT15] O. Klopp and A. B. Tsybakov. Estimation of matrices with row sparsity. *Problems of Information Transmission*, 51(4):335–348, 2015.

[KTV16] O. Klopp, A. B. Tsybakov, and N. Verzelen. Oracle inequalities for network models and sparse graphon estimation. *to appear in The Annals of Statistics*, 2016.

[Tch05] Olga Tchebotareva. An example of embedded singular continuous spectrum for one-dimensional Schrödinger operators. *Lett. Math. Phys.*, 72(3):225–231, 2005.

Articles in peer-reviewed conferences

[LKMS14] J. Lafond, O. Klopp, E. Moulines, and J. Salmon.. *Probabilistic low-rank matrix completion on finite alphabets.* Advances in Neural Information Processing Systems 27 (NIPS), (2014)

Preprint

[CKLN16] A. Carpentier, O. Klopp, M. Löffler, and R. Nickl. *Adaptive confidence sets for matrix completion*. *ArXiv e-prints*, August 2016.

All the papers are available on my website :

`http://kloppolga.perso.math.cnrs.fr/publi.html`

Bibliography

[1] Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Stat.*, 34(2):584–653, 2006.

[2] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 2012.

[3] Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[4] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[5] Peter J Bickel, Aiyou Chen, Elizaveta Levina, et al. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

[6] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

[7] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*, pages 1369–1377, 2015.

[8] F. Bunea, Y. She, and M. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*, 39:1282–1309, 2011.

[9] J. Cai, E. J Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[10] T. T. Cai and W-X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *JMLR*, 14:3619–3647, 2013.

[11] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010.

[12] T. Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electron. J. Stat.*, 10(1):1493–1525, 2016.

[13] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(1):1–37, 2009.

[14] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[15] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359, 2011.

[16] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[17] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010.

[18] Stanley H. Chan and Edoardo M. Airoldi. A consistent histogram estimator for exchangeable graph models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 208–216, 2014.

[19] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[20] S. Chatterjee. Matrix estimation by universal singular value thresholding. *Ann. Statist.*, 43(1):177–214, 02 2015.

[21] Yuand Chen, Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust matrix completion with corrupted columns. *ICML*, pages 873–880, 2011.

[22] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324 –4337, 2013.

[23] Mark A. Davenport, Yaniv Plan, Ewout van den Berg, and Mary Wootters. 1-bit matrix completion. *Inf. Inference*, 3(3):189–223, 2014.

[24] C. Dhanjal, R. Gaudel, and S. Clémençon. Online matrix completion through nuclear norm regularisation. *SIAM International Conference on Data Mining*, pages 623–631, 2014.

[25] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008.

[26] David L. Donoho and Iain M. Johnstone. Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, 99(2):277–303, 1994.

[27] M. Dudík, Z. Harchaoui, and J. Malick. Lifted coordinate descent for learning with trace-norm regularization. *AISTATS of JMLR Proceedings*, 22:327–336, 2012.

[28] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011. PMID: 22279246.

[29] R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. *Journal 24nd Annual Conference on Learning Theory (COLT)*, 2011.

[30] Rina Foygel, Ohad Shamir, Nati Srebro, and Ruslan R Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2133–2141. Curran Associates, Inc., 2011.

[31] Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

[32] Christophe Giraud. Low rank multivariate regression. *Electronic Journal of Statistics*, 5:775–799, 2011.

[33] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, 2011.

[34] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B*, 55(4):757–796, 1993. With discussion and a reply by the authors.

[35] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

[36] Il'dar Abdulovich Ibragimov and Rafail Zalmanovich Khasminski. *Statistical estimation : asymptotic theory*. Applications of mathematics. Springer, New York, Heidelberg, Berlin, 1981.

[37] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11:2057–2078, 2010.

[38] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[39] Xiaodong Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1), 2013.

[40] Heng Lian. Variable selection for high-dimensional generalized varying-coefficient models. *Statist. Sinica*, 22(4):1563–1588, 2012.

[41] Heng Lian and Shujie Ma. Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates. *arXiv preprint arXiv:1309.6058*, 2013.

[42] L. Lovasz and B. Szegedy. Limits of dense graph sequences. *ArXiv Mathematics e-prints*, August 2004.

[43] László Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2012.

[44] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322, 2010.

[45] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

[46] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, 2012.

[47] Mathew Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.

[48] Ben Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[49] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Stat.*, 39(2):731–771, 2011.

[50] Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012.

[51] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39 (2):887–930, 2011.

[52] Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *Ann. Statist.*, 41(3):1406–1430, 2013.

[53] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[54] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[55] E. Grosse W. S. Cleveland and W. M. Shyu. *Statistical Models in S*. Wadsworth and Brooks/Cole, 1992.

[56] Fengrong Wei, Jian Huang, and Hongzhe Li. Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica*, 21(4):1515–1540, 2011.

[57] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation. *arXiv preprint arXiv:1309.5936*, 2013.

[58] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Inform. Theory*, 58(5):3047–3064, 2012.

[59] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. *COLT*, 2014.