Tight Risk Bound for High Dimensional Time Series Completion

Pierre Alquier

RIKEN
AIP Center for Advanced Intelligence Project

EcoDep 2021 Conference
September 16, 2021

Co-authors

Alquier, P., Marie, N. and Rosier, A. (2021). Tight Risk Bound for High Dimensional Time Series Completion. *Preprint arXiv :2102.08178*.

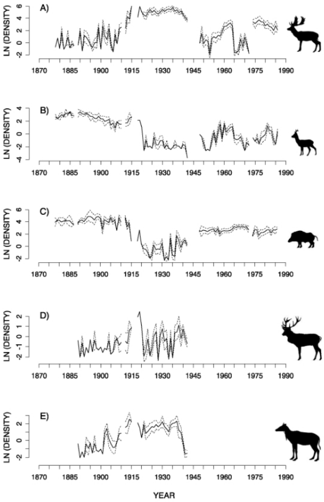
Nicolas Marie

Université Paris Nanterre

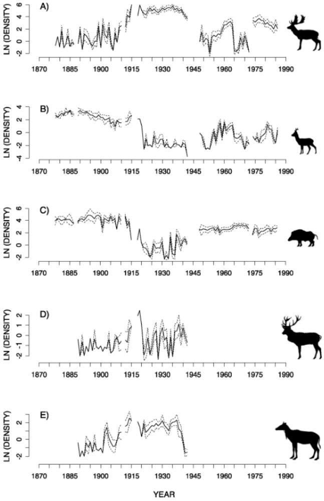
Amélie Rosier

ESME Sudria and Université Paris Nanterre

Multivariate time series



Multivariate time series



S. Imperio *et al.* (2010). Investigating population dynamics in ungulates : Do hunting statistics make up a good index of population abundance ? *Wildlife Biology*.

- multivariate series
- correlations
- noisy observations
- missing entries

Partially observed multivariate time series

i	...	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	...
1	...		12.5			17			...
2	...	1.2			3.8			2.9	...
3	...			0		7.2			...
4	...				4.2	3.1	2.4	2.3	...
5	...	23.1			45.1	39.9			...
6	...		4.1	4.1		6.3		2.9	...
7	...	0.1		0.9	0				...
8	...					34.7			...
⋮	⋮				⋮				⋱

Examples

- econometrics : panel data with missing entries,
- industry : data from sensors at multiple locations,
- ecology : spatial data with observations from a few sites only at each date,
- ...
- more generally, any situation where we have multivariate time series and each measurement is expensive.

Matrix completion methods

- matrix completion algorithms exist, and were successful in many applications.
- many of them are based on a low-rank assumption and on matrix factorization.
- however, the theory was developped only in the independent case.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 56, NO. 5, MAY 2010 2053

The Power of Convex Relaxation: Near-Optimal Matrix Completion

Emmanuel J. Candès, *Associate Member, IEEE*, and Terence Tao

***Abstract*—This paper is concerned with the problem of recovering an unknown matrix from a small fraction of its entries. This is known as the *matrix completion* problem, and comes up in a great number of applications, including the famous *Netflix Prize* and other similar questions in collaborative filtering. In general, accurate recovery of a matrix from a small number of entries is impossible, but the knowledge that the unknown matrix has low rank radically changes this premise, making the search for solutions meaningful. This paper presents optimality results quantifying the minimum number of entries needed to recover a matrix of rank r exactly by any method whatsoever (the information theoretic limit). More importantly, the paper shows that, under certain incoherence assumptions on the singular vectors of the matrix, recovery is possible by solving a convenient convex program as soon as the number of entries is on the order of the information theoretic limit (up to logarithmic factors). This convex program simply finds, among all matrices consistent with the observed entries, that with minimum nuclear norm. As an example, we show that on the order of $nr\log(n)$ samples are needed to recover a random $n \times n$ matrix of rank r by any method, and to be sure, nuclear norm minimization succeeds as soon as the number of entries is of the form $nr\text{polylog}(n)$.**

***Index Terms*—Duality in optimization, free probability, low-rank matrices, matrix completion, nuclear norm minimization, random matrices and techniques from random matrix theory, semidefinite programming.**

I. Introduction

A. Motivation

IMAGINE we have an $n_1 \times n_2$ array of real[1] numbers and that we are interested in knowing the value of each of the $n_1 n_2$ entries in this array. Suppose, however, that we only get to see a small number of the entries so that most of the elements about which we wish information are simply missing. Is it possible from the available entries to guess the many entries that we have not seen? This problem is now known as the *matrix completion* problem [7], and comes up in a great number of applications, including the famous *Netflix Prize* and other similar questions in collaborative filtering [12]. In a nutshell, collaborative filtering is the task of making automatic predictions about the interests of a user by collecting taste information from many users. Netflix is a commercial company implementing collaborative filtering, and seeks to predict users' movie preferences from just a few ratings per user. There are many other such recommendation systems proposed by Amazon, Barnes and Noble, and Apple Inc. to name just a few. In each instance, we have a partial list about a user's preferences for a few rated items, and would like to predict his/her preferences for all items from this and other information gleaned from many other users.

In mathematical terms, the problem may be posed as follows: we have a data matrix $M \in \mathbb{R}^{n_1 \times n_2}$ which we would like to know as precisely as possible. Unfortunately, the only information available about M is a sampled set of entries M_{ij}, $(i,j) \in \Omega$, where Ω is a subset of the complete set of entries $[n_1] \times [n_2]$ (here, and in the sequel, $[n]$ denotes the list $\{1, \ldots, n\}$). Clearly, this problem is ill-posed for there is no way to guess the missing entries without making any assumption about the matrix M.

An increasingly common assumption in the field is to suppose that the unknown matrix M has low rank or has approximately low rank. In a recommendation system, this makes sense because often times, only a few factors contribute to an individual's taste. In [7], the authors showed that this premise radically changes the problem, making the search for solutions meaningful. Before reviewing these results, we would like to emphasize that the problem of recovering a low-rank matrix from a sample of its entries, and by extension from fewer linear functionals about the matrix, comes up in many application areas other than collaborative filtering. For instance, the completion problem also arises in computer vision. There, many pixels may be missing in digital images because of occlusion or tracking failures in a video sequence. Recovering a scene and inferring camera motion from a sequence of images is a matrix completion problem known as the structure-from-motion problem [9], [24]. Other examples include system identification in control [20], multiclass learning in data analysis [1]–[3], global positioning—e.g., of sensors in a network—from partial distance information [5], [22], [23], remote sensing applications in signal processing where we would like to infer a full covariance matrix from partially observed correlations [26], and many statistical problems involving succinct factor models.

Manuscript received March 11, 2009; revised August 12, 2009. Current version published April 21, 2010. E. J. Candès was supported in part by ONR grants N00014-09-1-0469 and N00014-08-1-0749 and in part by the NSF Waterman Award. T. Tao was supported in part by a grant from the MacArthur Foundation, in part by NSF grant DMS-0649473, and in that part by the NSF Waterman Award.

E. J. Candès is with the Department of Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: emmanuel@acm.caltech.edu).

T. Tao is with the Department of Mathematics, University of California, Los Angeles, CA 90095 USA (e-mail: tao@math.ucla.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2010.2044061

[1]Much of the discussion below, as well as our main results, applies also to the case of complex matrix completion, with some minor adjustments in the absolute constants; but for simplicity we restrict attention to the real case.

Contents

Contents

Classical example : collaborative filtering

									
Stan			👍		👎		👍		
Pierre		👍	👍	👍	👍	👍	👍	👍	👍
Zoe	👍	👎					👍		
Bob			👍	👎				👎	
Oscar				👍		👍		👍	
Léa		👍	👎		👍				
Tony			👍	👎				👎	👍

A statistical model

There is a $d \times T$ matrix M and n i.i.d observations $Y_1, \dots, Y_n$ drawn as :

- (i_ℓ, j_ℓ) drawn uniformly on $\{1, \dots, d\} \times \{1, \dots, T\}$,
- $Y_\ell = M_{i_\ell, j_\ell} + \varepsilon_\ell$

where ε_ℓ is some noise ($= 0$ in the first papers on the topic, subgaussian with variance σ^2 later).

A statistical model

There is a $d \times T$ matrix M and n i.i.d observations $Y_1, \dots, Y_n$ drawn as :

- (i_ℓ, j_ℓ) drawn uniformly on $\{1, \dots, d\} \times \{1, \dots, T\}$,
- $Y_\ell = M_{i_\ell, j_\ell} + \varepsilon_\ell$

where ε_ℓ is some noise ($= 0$ in the first papers on the topic, subgaussian with variance σ^2 later).

Key assumption : $k := \mathrm{rank}(M) \ll \min(d, T) = K$.

SVD & matrix factorization

$$M = \underbrace{\left(\begin{array}{c|c|c|c} U_1 & \dots & U_k & \dots \end{array}\right)}_{=U\ (d \times K)} \underbrace{\begin{pmatrix} \sigma_1 & 0 & \dots & & & \\ 0 & \ddots & 0 & \dots & & \\ \vdots & & \sigma_k & & & \\ & & & 0 & & \\ & & & & \ddots & \end{pmatrix}}_{=\Sigma\ (K \times K)} \underbrace{\begin{pmatrix} V_1^T \\ \hline \vdots \\ \hline V_k^T \\ \hline \vdots \end{pmatrix}}_{=V^T\ (K \times T)}$$

SVD & matrix factorization

$$M = \underbrace{\left(\begin{array}{c|c|c|c} U_1 & \dots & U_k & \dots \end{array}\right)}_{=U\ (d\times K)} \underbrace{\begin{pmatrix} \sigma_1 & 0 & \dots & & \\ 0 & \ddots & 0 & \dots & \\ \vdots & & \sigma_k & & \\ & & & 0 & \\ & & & & \ddots \end{pmatrix}}_{=\Sigma\ (K\times K)} \underbrace{\begin{pmatrix} V_1^T \\ \hline \vdots \\ \hline V_k^T \\ \hline \vdots \end{pmatrix}}_{=V^T\ (K\times T)}$$

$$M = \underbrace{\left(\begin{array}{c|c|c} U_1 & \dots & U_k \end{array}\right) \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_k \end{pmatrix}}_{=A\ (d\times k)} \underbrace{\begin{pmatrix} V_1^T \\ \hline \vdots \\ \hline V_k^T \end{pmatrix}}_{=B\ (k\times T)}$$

Estimation

$$\hat{M}^{\lambda} = \arg\min_{X} \left\{ \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2 + \lambda \sum_{h=1}^{\min(d,T)} \sigma_h(X) \right\}.$$

Estimation

$$\hat{M}^{\lambda} = \arg\min_{X} \left\{ \sum_{\ell=1}^{n} (Y_{\ell} - X_{i_{\ell},j_{\ell}})^2 + \lambda \sum_{h=1}^{\min(d,T)} \sigma_h(X) \right\}.$$

Theorem

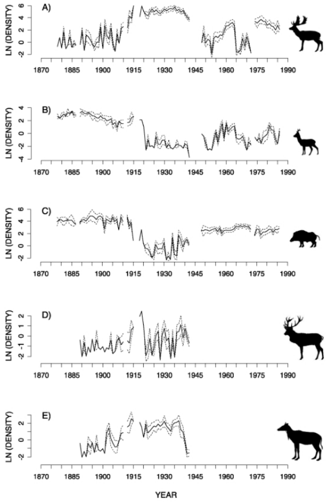
For a well chosen λ that does not depend on k, and under minimal assumptions on M, with large probability

$$\frac{1}{dT} \sum_{i,j} \left(\hat{M}^{\lambda}_{i,j} - M_{i,j} \right)^2 \leq \mathrm{Cst} \frac{\sigma k(d+T)\log(d+T)}{n}$$

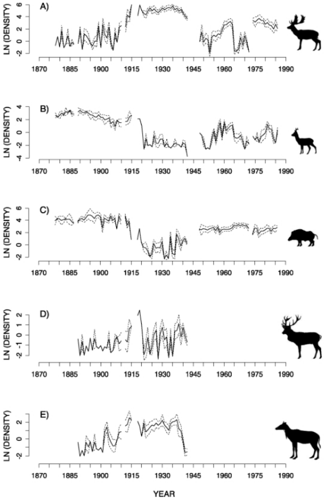
Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*.

Contents

Time series completion : the model



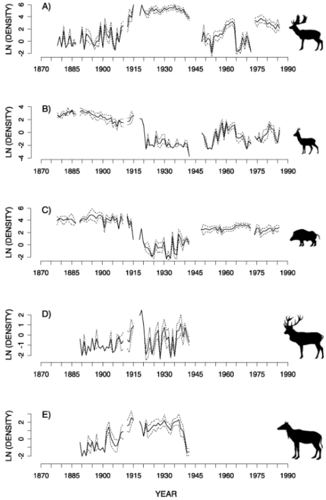
Time series completion : the model



- low-rank trend :

$$M = \underbrace{A}_{d \times k} \underbrace{B}_{k \times T}$$

Time series completion : the model



- low-rank trend :

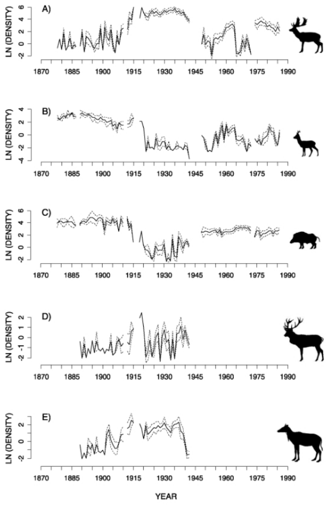
$$M = \underbrace{A}_{d \times k} \underbrace{B}_{k \times T}$$

- temporal correlated noise ε :

$$\varepsilon_{i,t} \text{ indep. } \varepsilon_{j,t'} \ (i \neq j)$$

$$\varepsilon_{i,t} \text{ not indep. } \varepsilon_{i,t'}$$

Time series completion : the model



- low-rank trend :

$$M = \underbrace{A}_{d\times k}\underbrace{B}_{k\times T}$$

- temporal correlated noise ε :

$$\varepsilon_{i,t} \text{ indep. } \varepsilon_{j,t'} \ (i \neq j)$$

$$\varepsilon_{i,t} \text{ not indep. } \varepsilon_{i,t'}$$

- (i_ℓ, t_ℓ) i.i.d uniform, ξ_ℓ observation noise :

$$Y_\ell = M_{i_\ell,t_\ell} + \varepsilon_{i_\ell,t_\ell} + \xi_\ell.$$

Assumptions

Reminder : the model

$$Y_\ell = M_{i_\ell,t_\ell} + \varepsilon_{i_\ell,t_\ell} + \xi_\ell.$$

Assumptions

Reminder : the model

$$Y_\ell = M_{i_\ell,t_\ell} + \varepsilon_{i_\ell,t_\ell} + \xi_\ell.$$

- $M = \underbrace{A}_{d\times k}\underbrace{B}_{k\times T}$ and $|A_{i,h}|, |B_{h,t}| \leq c_{A,B}/\sqrt{k}$.
- (i_ℓ, t_ℓ) i.i.d uniform on $\{1,\dots,d\}\times\{1\dots,T\}$;
- $(\varepsilon_{i,t})_{t=1,\dots,T}$ is a bounded, ϕ-mixing time series :

$$|\varepsilon_{i,t}| \leq m_\varepsilon \text{ and } \sum_{t=1}^{\infty}\phi_{\varepsilon_{i,\cdot}}(t) \leq \Phi_\varepsilon.$$

- (ξ_ℓ) are i.i.d, sub-exponential variables : for $k \geq 2$,

$$\mathbb{E}(|\xi_\ell|^q) \leq \frac{v_\xi c_\xi^{q-2} q!}{2}.$$

Estimator and risk bound

$$\hat{M}^{(k)} = \operatorname*{arg\,min}_{\underbrace{X}_{d\times T} = \underbrace{A}_{d\times k}\underbrace{B}_{k\times T}} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2.$$

Estimator and risk bound

$$\hat{M}^{(k)} = \underset{\underbrace{X}_{d\times T} = \underbrace{A}_{d\times k}\underbrace{B}_{k\times T}}{\arg\min} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2.$$

Theorem

With probability at least $1-\eta$,

$$\frac{1}{dT}\sum_{i,j}\left(\hat{M}^{(k)}_{i,j} - M_{i,j}\right)^2 \leq C\frac{k(d+T)\log(n) + \log\left(\frac{1}{\eta}\right)}{n}$$

where $C = C(c_{A,B}, m_\varepsilon, \Phi_\varepsilon, v_\xi, c_\xi)$ is known.

Remarks on the proof

1. decompose the difference between *empirical risk* and *expected risk* $\frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - X_{i_\ell,j_\ell})^2 - \frac{1}{dT}\sum_{i,j}(M_{i,j} - X_{i,j})^2$ in elementary terms.
2. some of these terms are sums of i.i.d variables. Bound them via Bernstein inequality. Some are sums of ϕ-mixing variables, use :

> Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*.

3. union bound.

Remarks on the proof

1. decompose the difference between *empirical risk* and *expected risk* $\frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - X_{i_\ell,j_\ell})^2 - \frac{1}{dT}\sum_{i,j}(M_{i,j} - X_{i,j})^2$ in elementary terms.
2. some of these terms are sums of i.i.d variables. Bound them via Bernstein inequality. Some are sums of ϕ-mixing variables, use :

Samson, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *The Annals of Probability*.

3. union bound.

REMARK : if the $\varepsilon_{i,\cdot}$ satisfy another notion of mixing or weak-dependence, we can use alternative versions of Bernstein inequality but this lead to slower rates of convergence, in $1/\sqrt{n}$.

Rank selection

$$\hat{k} = \arg\min_{1 \leq k \leq K} \left\{ \frac{1}{n} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2 + C' \frac{k(d+T)\log(n)}{n} \right\}$$

where $C' = C'(c_{A,B}, m_\varepsilon, \Phi_\varepsilon, v_\xi, c_\xi)$ is **known** but too large.

Rank selection

$$\hat{k} = \arg\min_{1\leq k\leq K} \left\{ \frac{1}{n}\sum_{\ell=1}^{n}(Y_\ell - X_{i_\ell,j_\ell})^2 + C' \frac{k(d+T)\log(n)}{n} \right\}$$

where $C' = C'(c_{A,B}, m_\varepsilon, \Phi_\varepsilon, v_\xi, c_\xi)$ is **known** but too large.

In practice : we use the slope heuristic to calibrate a better C'.

Rank selection

$$\hat{k} = \arg\min_{1\leq k\leq K} \left\{ \frac{1}{n}\sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell,j_\ell})^2 + C' \frac{k(d+T)\log(n)}{n} \right\}$$

where $C' = C'(c_{A,B}, m_\varepsilon, \Phi_\varepsilon, v_\xi, c_\xi)$ is **known** but too large.

In practice : we use the slope heuristic to calibrate a better C'.

Theorem

With probability at least $1-\eta$,

$$\frac{1}{dT}\sum_{i,j}\left(\hat{M}^{(\hat{k})}_{i,j} - M_{i,j}\right)^2 \leq C'' \frac{k(d+T)\log(n) + \log\left(\frac{1}{\eta}\right)}{n}.$$

Time series with a structure

Example : assume that the trends in M are p-periodic. This means that

$$\underbrace{M}_{d\times T} = \underbrace{C}_{d\times p} \underbrace{(I_p|\dots|I_p)}_{=\Lambda\ (p\times T)}.$$

Time series with a structure

Example : assume that the trends in M are p-periodic. This means that

$$\underbrace{M}_{d\times T} = \underbrace{C}_{d\times p} \underbrace{(I_p| \dots |I_p)}_{=\Lambda \ (p\times T)}.$$

More generally, we can assume that there is a known structure in M :

$$\underbrace{M}_{d\times T} = \underbrace{C}_{d\times p} \underbrace{\Lambda}_{p\times T}$$

and still add the initial "low-rank decomposition" to ensure correlations in the rows :

$$\underbrace{M}_{d\times T} = \underbrace{A}_{d\times k} \underbrace{B}_{k\times p} \underbrace{\Lambda}_{p\times T}.$$

Faster rates

$$\hat{M}^{(k)} = \underset{\underbrace{X}_{d\times T} = \underbrace{A}_{d\times k}\underbrace{B}_{k\times p}\underbrace{\Lambda}_{p\times T}}{\arg\min} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2.$$

Faster rates

$$\hat{M}^{(k)} = \mathop{\arg\min}_{\underbrace{X}_{d\times T} = \underbrace{A}_{d\times k}\underbrace{B}_{k\times p}\underbrace{\Lambda}_{p\times T}} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2.$$

Theorem

With probability at least $1-\eta$,

$$\frac{1}{dT}\sum_{i,j}\left(\hat{M}^{(k)}_{i,j} - M_{i,j}\right)^2 \leq C \frac{k(d+p)\log(n) + \log\left(\frac{1}{\eta}\right)}{n}.$$

Faster rates

$$\hat{M}^{(k)} = \underset{\underbrace{X}_{d\times T} = \underbrace{A}_{d\times k}\underbrace{B}_{k\times p}\underbrace{\Lambda}_{p\times T}}{\arg\min} \sum_{\ell=1}^{n} (Y_\ell - X_{i_\ell, j_\ell})^2.$$

Theorem

With probability at least $1-\eta$,

$$\frac{1}{dT}\sum_{i,j}\left(\hat{M}^{(k)}_{i,j} - M_{i,j}\right)^2 \leq C \frac{k(d+p)\log(n) + \log\left(\frac{1}{\eta}\right)}{n}.$$

We also have a similar rank-selection procedure.

RIKEN AIP : position in the ABI team

RIKEN AIP
Center for Advanced Intelligence Project

Approximate Bayesian Inference team (ABI), lead by Emtiyaz Khan

Please visit the team website

https ://team-approx-bayes.github.io/

Open Position : Research Scientist (1 position, Female only)

- research only (= chargé de recherches),
- indefinite-term,
- located in Tokyo center.