COLUMN SUBSET SELECTION AND APPLICATIONS TO COMPESSED SENSING AND FEATURE EXTRACTION

Stéphane Chrétien*
National Physical Laboratory, Teddington, UK - stephane.chretien@npl.co.uk

Zhen Wai Olivier Ho
Université de Bourgogne Franche Comté, Besançon, France - zhen_wai_olivier.ho@univ-fcomte.fr

Abstract The column selection problem lies at the crossroad of many applications in mathematics, statistics and machine learning. In this paper, we survey some of the key results in this field. We also provide a new perturbation result which leads to a simple and efficient greedy column selection method.

.

Keywords: Eigenvalue perturbation; feature extraction; incoherence.

1. Introduction Let X in $R^{n\times p}$ be a matrix such that all columns of X have unit euclidean ℓ_2-norm and for any index subset $T \subset \{1,\dots,p\}$, let X_T denote the submatrix of X obtained by extracting the columns of X indexed by T. The problem of extracting a subset of columns of X such that the resulting matrix is well conditioned has been studied for quite some time both in the pure and in the applied mathematics literature. In statistics, this problem is usually addressed in the context where more variables are available than observations and a relevant set of covariates has to be chosen to stably represent all the variables at hand. In numerical analysis, one usually speaks of Rank Revealing QR factorization methods. In Computational Geometry the problem of finding the j-simplex of maximal volume has been investigated using column selection. In functional analysis, one of the most famous problems of this type is the Restricted Invertibility Problem of Bourgain and Tzafriri. Several selection criteria have been investigated depending on the application. In the Restricted Invertibility Problem, one is concerned with finding the largest number of columns such that the resulting submatrix has its smallest singular value larger than or equal to a given constant. In Rank Revealing QR Factorization, much interest usually goes into controlling the extreme singular values as well as the spacings between them. In Computational Geometry, one is interested in selecting a subset of columns from X with largest possible determinant. From a computational viewpoint, choosing the columns so as to optimize certain criteria associated with well conditioning of the resulting submatrix is often NP-hard. Greedy algorithm with good approximation guaranties are available. Various clever other techniques have been used for the practical solution of the column selection problems. Among them, randomized algorithms have been very popular. A deterministic algorithm was proposed by Spielmann and Srivastava for the Restricted Invertibility Problem. This method was then further extended to the problem of controlling the smallest and the largest singular values at the same time by Youssef. In the numerical analysis community several authors have come up with very interesting methods as well.
The goal of this paper is to give an overview of the columns selection problem and its various applications in high dimensional statistics. We will also present a very simple deterministic method which controls the extreme singular values when the matrix has some intrinsic incoherence. Applications to feature extraction will be described. In particular, we will show that our method can be successfully used in practice on some difficult machine learning problems.

2. Previous results.
2.1 The restricted invertibility problem. For the Restricted Invertibility problem, Bourgain and Tzafriri obtained the following result for square matrices:

Theorem (A) *[Bourgain-Tzafriri, '87] Given a $p \times p$ matrix X whose columns have unit ℓ_2-norm, there exists $T \subset \{1, \dots, p\}$ with $|T| \geq d\dfrac{p}{\|X\|^2}$ such that $C \leq \lambda_{\min}(X_T^t X_T)$, where d and C are absolute constants.*

See also (Tropp, '08) for a simpler proof. Vershynin (Vershynin 01) generalized Bourgain and Tzafriri's result to the case of rectangular matrices and the estimate of $|T|$ was improved as follows.

Theorem (B) *[Vershynin, '01] Given a $n \times p$ matrix X and letting $\widetilde{X}$ be the matrix obtained from X by ℓ_2-normalizing its columns. Then, for any $\varepsilon \in (0,1)$, there exists $T \subset \{1, \dots, p\}$ with*

$$|T| \geq (1-\varepsilon)\frac{\|X\|_{HS}^2}{\|X\|^2}$$

such that $C_1(\varepsilon) \leq \lambda_{\min}(\widetilde{X}_T^t \widetilde{X}_T) \leq \lambda_{\max}(\widetilde{X}_T^t \widetilde{X}_T) \leq C_2(\varepsilon)$.

Recently, Spielman and Srivastava proposed in (Spielman and Srivastava, '12) a deterministic construction of T which allows them to obtain the following result.

Theorem (C) *[Spielman-Srivastava, '12] Let X be a $p \times p$ matrix and $\varepsilon \in (0,1)$. Then there exists $T \subset \{1, \dots, p\}$ with $|T| \geq (1-\varepsilon)^2 \dfrac{\|X\|_{HS}^2}{\|X\|^2}$ such that $\varepsilon^2 \dfrac{\|X\|^2}{p} \leq \lambda_{\min}(X_T^t X_T)$.*

The technique of proof relies on new constructions and inequalities which are thoroughly explained in the Bourbaki seminar of Naor (Naor, '12). Using these techniques, Youssef (Youssef, '13) improved Vershynin's result as:

Theorem (D) *[Youssef, '13] Given a $n \times p$ matrix X and letting $\widetilde{X}$ be the matrix obtained from X by ℓ_2-normalizing its columns. Then, for any $\varepsilon \in (0,1)$, there exists $T \subset \{1, \dots, p\}$ with $|T| \geq \dfrac{\varepsilon^2}{9}\dfrac{\|X\|_{HS}^2}{\|X\|^2}$ such that $1-\varepsilon \leq \lambda_{\min}(\widetilde{X}_T^t \widetilde{X}_T) \leq \lambda_{\max}(\widetilde{X}_T^t \widetilde{X}_T) \leq 1+\varepsilon$.*

2.2 Other criteria for column selection. Other techniques for columns extraction are available using different criteria. In what follows, we list some of the existing results in this direction.
The paper (deHoog-Mattheij, '07) studies the problem of selecting a subset of k columns from X such that the pseudo-inverse of the sampled matrix has as small a norm as possible. Their approach is greedy and deterministic. The idea is to proceed by removing one column at a time from X. In the first iteration of the algorithm, they remove the column with index i_1, where

$$i_1 = \text{argmin trace}\left(\left(X - x_i x_i^t\right)^{-1}\right).$$

Then, the column i_1 is removed from X and the resulting matrix is denoted by X_1. Then, i_2 is chosen as

$$i_2 = \text{argmin trace}\left(\left(X_1 - x_i x_i^t\right)^{-1}\right),$$

and so on and so forth in the same way for i_3, i_4, etc. The following theorem holds for this procedure when the number of extracted columns is larger than n.

Theorem (E) *[deHoog-Mattheij, '07] If X is such that the removal of a single column does not result in a rank deficient matrix, we have*

$$\|X_T^\dagger\|_F^2 \leq \frac{p-n+1}{|T|-n+1}\|X_T^\dagger\|_F^2. \tag{0.1}$$

The work of Gu and Eisenstat (Gu-Eisenstat, '95) about rank revealing factorisation is also widely cited in this area of research. Without entering into the details, their algorithm provides a numerically stable way to compute a subset T of cardinality $|T| \leq n$ with bounds on all the nonzero singular values of the type

$$\sigma_i^2(X_t) \geq \frac{1}{1 + f^2 n(p-n)} \, \sigma_i^2(X) \tag{0.2}$$

$i = 1, \ldots, n$ and a certain parameter f.
A greedy approach was also proposed in (Boutsidis-Drineas-Magdon-Ismail, '11) which was proved to be nearly optimal. In (Avron-Boutsidis, '13), a new random sampling based method is proposed for this problem with many possible norms. They are able to devise a clever method which gives the same results as (deHoog-Mattheij, '07) but systematically avoids selecting a submatrix which lowers the rank. They also establish the remarkable fact that the combinatorial problem of finding a low-stretch spanning tree in an undirected graph corresponds to this same kind of subset selection problem although the column selection approach is not optimal for this problem from a computational complexity perspective.
Convex optimisation was also put at work for the columns selection problem as in (Joshi-Boyd, '09) using a clever relaxation and a D-optimality type criterion. Tropp (Tropp, '09) also addressed the problem using convex optimisation and in particular Semi-Definite Programming solution to the Pietch and Grothendieck factorisations. In the same spirit, column selection based on maximum volume type criteria have also been extensively studied, such as in (Nikolov, '15) where an efficient randomised algorithm is described.
3 Results based on the coherence and application to Compressed Sensing. Incoherent subset selection is also a very interesting class of problems of paramount interest in Compressed Sensing and high dimensional regression.
3.1 Random sampling. In these problems, one usually want to know whether most submatrices with s columns extracted from a given X are r_0-quasi isometry when T is a random index subset of size s of $\{1, \ldots, p\}$ drawn uniformly at random and X_T is the matrix obtained by extracting the columns of X indexed by T. By an r_0-quasi isometry, we simply mean $\|X_T^t X_T - \mathrm{Id}\| \leq r_0$. In the sequel, we assume that the columns of X have unit norm.
The uniform version of the quasi-isometry property, i.e. satisfied for all possible T's, is called the Restricted Isometry Property (RIP) and has been widely studied for random i.i.d. subgaussian matrices. Recent works such as (Candès-Plan, '09) proved that the quasi isometry property holds with high probability for matrices satisfying an certain incoherence assumption. Checking that a matrix is sufficiently incoherent is easy to check in practice. Such types of result are therefore of great potential interest for a wide class of problems involving high dimensional linear or nonlinear regression models.
In a recent work based on the landmark paper (Rudelson, '99), Tropp proved the following theorem.

Theorem (F) *[Tropp, '08] Let A be an $n \times n$ Hermitian matrix, decomposed into diagonal and off-diagonal parts: $A = D + H$. Fix p in $[2, +\infty)$, and set $q = \max\{p, 2\log(n)\}$. Then*

$$\mathbb{E}_p \|RAR\| \leq C \left[q \mathbb{E}_p \|RHR\|_{\max} + \sqrt{\delta q} \mathbb{E}_p \|HR\|_{1,2} + \delta \|H\| \right] + \mathbb{E}_p \|RDR\| .$$

Here, R denotes the square diagonal "selector" matrix whose j^{th} diagonal entry is δ_j, where $\{\delta_j\}$ denotes a sequence of independent Bernoulli 0–1 random variables with common expectation δ, and the symbol $\mathbb{E}_p$ denotes the L_p norm $(\mathbb{E}|\cdot|^p)^{1/p}$. The proof heavily relies on the Non-Commutative Kintchin inequality.
Using this result and Markov's inequality, Candès and Plan proved that the 1/2-quasi isometry property holds with probability greater than $1 - p^{-2\log(2)}$ when $s \leq p/(4\|X\|^2)$ and the coherence of X, i.e. $\max |X_k^t X_l|$, $k \neq l$, is sufficiently small. The r_0-quasi isometry property then holds with high probability under easily checkable assumptions on X. Using refined tail decoupling techniques, we recently improved this last result in the following way.

Theorem (G) *[Chrétien-Darses, 13] Let $r \in (0,1)$, $\alpha \geq 1$. Let us be given a full rank matrix $X \in \mathbb{R}^{n \times p}$ and a positive integer s, such that*

$$\mu(X) \leq \frac{r}{(1+\alpha)\log p} \tag{0.3}$$

$$s \leq \frac{r^2}{(1+\alpha)e^2} \, \frac{p}{\|X\|^2 \log p}. \tag{0.4}$$

Let $T \subset \{1, \dots, p\}$ be a random support with uniform distribution on index sets with cardinal s. Then the following bound holds:

$$\mathbb{P}\left(\|X_T^t X_T - \mathrm{Id}_s\| \geq r\right) \quad \leq \quad \frac{1944}{p^\alpha}. \tag{0.5}$$

3.2 Gittens' result on Nyström extensions. Nyström extensions are a class of algorithms whose objective is to find a low-rank approximations to positive semidefinite (PSD) matrices by sampling from their columns. (Gittens, '11) considers a special case denoted the "naïve Nyström extension" in which the columns are sampled uniformly without replacement.

Theorem (H) *[Gittens, '11] Let A be a PSD matrix of size p (think of X^tX). Given an integer $k \leq n$, partition the eigenvalue decomposition of A as follows*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} U_1^t \\ U_2^t \end{bmatrix} \tag{0.6}$$

with $U_1 \in \mathbb{R}^{p \times k}$, $U_2 \in \mathbb{R}^{p \times (p-k)}$. Let τ denote the coherence of U_1. For any $\varepsilon \in (0,1)$, if T is chosen uniformly at random among subsets of $\{1, \dots, p\}$ with cardinality s with

$$s \geq \frac{2\tau k \log(k/\delta)}{(1-\varepsilon)^2}, \tag{0.7}$$

then the approximation error satisfies

$$\|A - AS_T(S_T^t A S_T)^\dagger S_T^t A\|_F \leq \lambda_{k+1}(A)\left(1 + \frac{p}{\varepsilon l}\right), \tag{0.8}$$

with probability larger than $1 - \delta$.

3. A new result on column subset selection for incoherent matrices

3.1 New perturbation results

Using the coherence as a way to choose the next column to select may be a natural way in a greedy approach. In order to study this type of procedure, we proved an instrumental general perturbation theorem. More precisely, if we consider a subset T_0 of $\{1, \dots, p\}$ and a submatrix X_{T_0} of X, the problem of studying the eigenvalue perturbations resulting from appending a column X_j to X_{T_0}, with $j \notin T_0$ can be studied using Cauchy's Interlacing Lemma. Based on this approach, we obtained the following result.

Theorem (I) *[Chretien-Ho, '17] Let $T_0 \subset \{1, \dots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geqslant \dots \geqslant \lambda_{s_0}$ be the eigenvalues of $X_{T_0} X_{T_0}^t$. Assume that for some $\alpha \in (0,1)$,*

$$\|X_{T_0}^t v\|_2^2 \leq \alpha\ s_0 \mu^2. \tag{0.9}$$

We have

$$\lambda_{s_0+1}\left(X_{T_0} X_{T_0}^t + X_j X_j^t\right) \geq \lambda_{s_0} - \frac{\alpha\ s_0 \mu^2}{1 - \lambda_{s_0}}. \tag{0.10}$$

If we append s_1 columns successively to the matrix X_{T_0}, we obtain the following result.

Theorem (J) *[Chretien-Ho, '17] Let $T_0 \subset \{1, \dots, p\}$ with $|T_0| = s_0$ and X_{T_0} a submatrix of X. Let $\lambda_1 \geqslant \dots \geqslant \lambda_{s_0}$ be the eigenvalues of $X_{T_0} X_{T_0}^t$. Let $T_1 \subset \{1, \dots, p\}$ with $|T_1| = s_1$ and $T_0 \cap T_1 = \emptyset$. Let*

$$\varepsilon_{min} = \min\left(\sqrt{\alpha\mu^2} \sum_{i=s_0}^{s_0+s_1} \sqrt{i}, \frac{\alpha\mu^2 s_0}{1-\lambda_{s_0}} + \frac{2(1-\lambda_{s_0})}{s_0} \sum_{i=s_0+1}^{s_0+s} \frac{i}{i-1}\right).$$

Then

$$\lambda_{s_0+s_1}\left(X_{T_0 \cup T_1}^t X_{T_0 \cup T_1}\right) \geq \lambda_{s_0} - \varepsilon_{min} \tag{0.11}$$

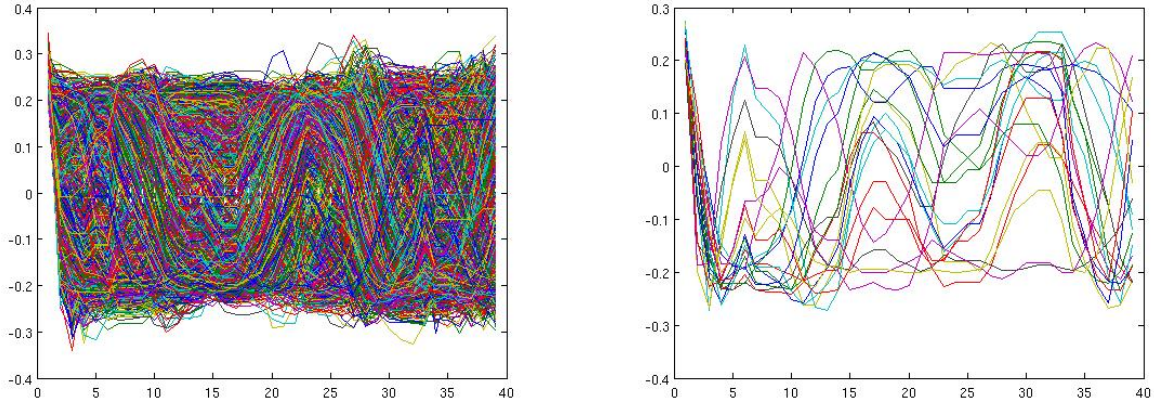

Figure 1: Time series (left) and 20 first times series discovered by the greedy approach (right)

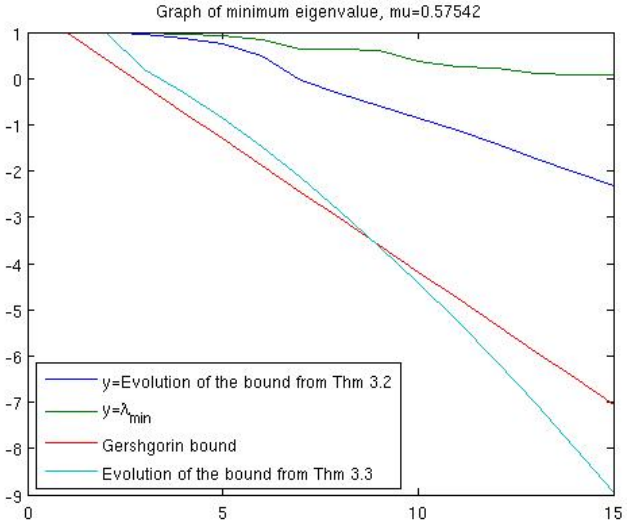

Figure 2: Extracting a submatrix sequentially by greedy column selection

3.2 Computer experiments

In this section, we show a particular computer experiment with a set of 10000 time series. The time series are shown in Figure 1. Our goal is to find a subset set of times series which represent the whole data set accurately. For this purpose, we use the bound in Theorem (I) in order to choose the next time series in a greedy fashion by minimising $\|X_T^t X_j\|_2$, $j \notin T$ at each step.
As Figure 2 shows that our bound significantly improves on the Gershgorin bound. Moreover, the graphs show that one could stop after 7 steps which implies that the times series could be clustered efficiently using 7 clusters.

4. Conclusions. In this paper, we surveyed the current results on the problem of column selection, an important problem in machine leaning and statistics for feature extraction, analysis of compressed sensing and the LASSO, numerical stabilisation, etc. We also showed that our bounds could lead to a very simple greedy method for clustering or feature extraction.

References

Avron, H. & Boutsidis, C. (2013). *Faster subset selection for matrices and applications. SIAM Journal on Matrix Analysis and Applications, 34(4), 1464-1499.*

C. Boutsidis, P. Drineas, and M. Magdon-Ismail, *Near optimal column based matrix reconstruction, in Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS), IEEE, Los Alamitos, CA, 2011, pp. 305-314.*

Bourgain, J. & Tzafriri, L. (1987). *Invertibility of 'large'submatrices with applications to the geometry of Banach spaces and harmonic analysis. Israel journal of mathematics, 57(2), 137-224.*

Candès, E. J., & Plan, Y. (2009). *Near-ideal model selection by âĎŞ1 minimization. The Annals of Statistics, 37(5A), 2145-2177.*

Chrétien, S., & Darses, S. (2012). *Invertibility of random submatrices via tail-decoupling and a matrix Chernoff inequality. Statistics Probability Letters, 82(7), 1479-1487.*

Chrétien, S., & Ho, Z.W.O., (2017), *Perturbation of incoherent matrices, in preparation.*

Deshpande, A. & Rademacher, L. (2010). *Efficient volume sampling for row/column subset selection. In 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA, pages 329-338. IEEE Computer Society.*

Gittens, A. (2011). *The spectral norm error of the naive Nystrom extension. arXiv preprint arXiv:1110.5305.*

Gu, M. & Eisenstat, S. C. (1995). *Downdating the singular value decomposition, SIAM J. Matrix Anal. Appl., 16, pp. 793-810.*

F. R. de Hoog and R. M. M. Mattheij, *Subset selection for matrices, Linear Algebra Appl., 422 (2007), pp. 349-359.*

Joshi, S. & Boyd, S. (2009). *Sensor selection via convex optimization. IEEE Transactions on Signal Processing, 57(2), 451-462.*

Naor, A. (2012). *Sparse quadratic forms and their geometric applications [following Batson, Spielman and Srivastava]. Astérisque.*

Nikolov, A. (2015). *Randomized rounding for the largest simplex problem. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (pp. 861-870). ACM.*

Spielman, D. A. & Srivastava, N. (2012). *An elementary proof of the restricted invertibility theorem. Israel Journal of Mathematics, 190(1), 83-91.*

Tropp, J. A. (2006). *The random paving property for uniformly bounded matrices, Studia Math, 185(1), 67-82.*

Tropp, J. A. (2009). *Column subset selection, matrix factorization, and eigenvalue optimization. In Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (pp. 978-986). Society for Industrial and Applied Mathematics.*

Vershynin, R. (2001). *John's decompositions: selecting a large part. Israel Journal of Mathematics, 122(1), 253-277.*

Youssef, P. (2013). *A note on column subset selection. International Mathematics Research Notices, rnt172.*