On a new method for controlling the entire spectrum in the problem of column subset selection

Stephane Chretien, Sebastien Darses

On a new method for controlling the entire spectrum in the problem of column subset selection

Article *in* Expositiones Mathematicae · April 2019

DOI: 10.1016/j.exmath.2019.02.002

2 authors, including:

Stéphane Chrétien
Université Lumiere Lyon 2
103 PUBLICATIONS **454** CITATIONS

SEE PROFILE

On a new method for controlling the entire spectrum in the problem of column subset selection

Stéphane Chrétien[a], Sébastien Darses[b,*]

[a] *National Physical Laboratory, Hampton Road, Teddington TW11 0LW, UK*
[b] *Aix-Marseille Université, Technopôle Château-Gombert 39, rue Frédéric Joliot-Curie 13453 Marseille Cedex 13*

Abstract

The problem of extracting a well conditioned submatrix from any rectangular matrix (with normalized columns) has been studied for some time in functional and harmonic analysis. In the seminal work of Bourgain and Tzafriri and many subsequent improvements, methods using random column selection were considered. Constructive approaches have been proposed lately, mainly sparked by the work of Batson, Spielman and Srivastava. The column selection problem we consider in this paper is concerned with extracting a well conditioned submatrix, and more precisely, a matrix with all its singular values being contained in the interval $[1-\varepsilon, 1+\varepsilon]$. Such results are known to have far reaching connections with many fields in mathematics and engineering.

Our main contribution is a new deterministic method that achieves the same order R for the number of selected columns as in Bourgain and Tzafriri's original Theorem, up to a $\log(R)$ multiplicative factor. Our analysis is elementary and shows how a simple eigenvalue perturbation argument can lead to an intuitive and very short proof. We also obtain individual lower and upper bounds for *each* singular value of the extracted matrix.

Keywords: Restricted Invertibility, Bourgain-Tzafriri Theorem, Column Subset Selection, Eigenvalue Perturbation.
2000 MSC: 68P, 68W, 42A.

1. Introduction

Let $X \in \mathbb{R}^{n\times p}$ be a matrix such that all columns of X have unit euclidean ℓ_2-norm. The problem of well conditioned column selection that we consider consists of finding the largest subset of columns of X such that the corresponding submatrix has all singular values in a prescribed interval $[1-\varepsilon, 1+\varepsilon]$. The one-sided problem of finding the largest

*Corresponding author
Email addresses: stephane.chretien@npl.co.uk (Stéphane Chrétien), sebastien.darses@univ-amu.fr (Sébastien Darses)

possible T such that $\lambda_{\min}(X_T^t X_T) \geq 1 - \varepsilon$ is called the Restricted Invertibility Problem and has a long history starting with the seminal work of Bourgain and Tzafriri [2].

Bourgain and Tzafriri's result has far reaching connections with many areas in mathematics and engineering and a renewed interest in Bourgain and Tzafriri's result was sparked by the introduction of new methods based on finite random matrix theory, followed by new deterministic methods and at the same time, by its strong similarity with RIP-type results in signal processing [8] and machine learning (under the name of "feature extraction") [1]. Recent contributions have shown that Column Selection may be an important ingredient in differential privacy [4], and discrepency minimisation [5]. Let us also recall that the original result of Bourgain and Tzafriri were motivated by application to harmonic analysis and the Kadison-Singer problem in operator theory [2].

2. Historical background

2.1. Bourgain and Tzafriri's original result

Bourgain and Tzfriri's original result on the Restricted Invertibility Problem can be stated as follows [2].

Theorem 2.1 ([2])**.** *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 $\lambda_{\min}(X_T^t X_T) \geq C$, where d and C are absolute constants.*

See also [7] for a simpler proof.

In [2], an application is given to harmonic analysis: let T be the circle with normalised Lebesgue measure ν and B be a subset of T with positive measure. Define the two norms

$$\|f\|_{L_2(B)} = \sqrt{\frac{1}{\nu(B)} \int_B f^2 d\nu} \quad \text{and} \quad \|f\|_{L_2(T)} = \sqrt{\int_T f^2 d\nu}. \tag{2.1}$$

Now suppose that the Fourier transform $\hat{f}$ of f is supported on a subset Λ of $\mathbb{Z}$. How dense can Λ be while still ensuring that the two norms are equivalent? Using Theorem 2.1, Bourgain and Tzafriri proved that there exists such a set Λ with density $c\nu(B)$ for which

$$\|f\|_{L^2(B)} \geq c \, \|f\|_{L^2(T)}. \tag{2.2}$$

2.2. Vershynin's generalization

Vershynin [10] generalized Bourgain and Tzafriri's result to the case of rectangular matrices and the estimate of $|T|$ was improved as follows.

Theorem 2.2 ([10])**.** *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 $C_1(\varepsilon) \leq \lambda_{\min}(\widetilde{X}_T^t \widetilde{X}_T) \leq \lambda_{\max}(\widetilde{X}_T^t \widetilde{X}_T) \leq C_2(\varepsilon)$.

Vershynin also presents in [9] an interesting application to communication systems. In that application, one assumes that a signal $x \in \mathbb{R}^n$ is encoded into a frame as follows:

$$x = \sum_{j=1}^{p} \langle x_j, x \rangle x_j. \tag{2.3}$$

Recall that the family x_j, $j = 1, \ldots, p$ is a frame if $\sum_{j=1}^{p} x_j x_j^t = I$. Of course, if $p > n$, the information contained in the coefficients $\langle x_j, x \rangle$, $j = 1, \ldots, p$ is redundant. This can be leveraged in communication systems when some coefficients can be lost during the transmission process. Theorem 5.1 in [9] determines a threshold for the probability of loosing a component at random under which one can still nearly recover the original signal.

2.3. Spielman and Srivastava's and Youssef's contributions

In [6], Spielman and Srivastava proposed in a *deterministic* construction of T which allows them to obtain the following result.

Theorem 2.3 ([6])**.** *Let X be a $p \times p$ matrix and $\varepsilon \in (0,1)$. Then there exists $T \subset \{1, \ldots, p\}$ with $|T| \geq (1-\varepsilon)^2 \frac{\|X\|_{HS}^2}{\|X\|^2}$ such that $\varepsilon^2 \frac{\|X\|^2}{p} \leq \lambda_{\min}(X_T^t X_T)$.*

The technique of proof relies on new constructions and inequalities which are thoroughly explained in Naor's Bourbaki seminar [3].

Using these techniques, Youssef [11] improved Vershynin's result as:

Theorem 2.4 ([11])**.** *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 $T \subset \{1, \ldots, p\}$ with $|T| \geq \frac{\varepsilon^2}{9} \frac{\|X\|_{HS}^2}{\|X\|^2}$ such that $1 - \varepsilon \leq \lambda_{\min}(\widetilde{X}_T^t \widetilde{X}_T) \leq \lambda_{\max}(\widetilde{X}_T^t \widetilde{X}_T) \leq 1 + \varepsilon$.*

3. A new elementary approach

3.1. Our contribution

Our main contribution is a short and elementary proof of the following result:

Theorem 3.1. *Given a $n \times p$ matrix X whose columns have unit ℓ_2-norm, a constant $\varepsilon \in (0,1)$ there exists $T \subset \{1, \ldots, p\}$ with $|T| \geq R$ and*

$$R \log R \;\; \leq \;\; \frac{\varepsilon^2}{4(1+\varepsilon)} \frac{p}{\|X\|^2}, \tag{3.4}$$

such that $1 - \varepsilon \leq \lambda_{\min}(X_T^t X_T) \leq \lambda_{\max}(X_T^t X_T) \leq 1 + \varepsilon$. Moreover, we have the following bounds on each of the individual eigenvalues:

$$1 - \sqrt{\frac{(1+\varepsilon)\|X\|^2 \log R}{p}} \; \frac{R + k - 1}{\sqrt{R}} \;\; \leq \;\; \lambda_k(X_T^t X_T) \;\; \leq \;\; 1 + \sqrt{\frac{(1+\varepsilon)\|X\|^2 \log R}{p}} \; \frac{2R - k}{\sqrt{R}}$$

for all $k = 1, \ldots, R$.

Notice that we loose a $\log(R)$ factor in (3.4) compared to the original result of Bourgain and Tzafriri. One the other hand, our method of proof is able to provide an individual control of each eigenvalue, which might be of independent interest. In data science in particular, it is intersting to know that either not all eigenvalues of the selected sub-matrix (of so called "features") achieve the worst case bound but may instead be more evenly distributed inside the interval.

4. Proof of Theorem 3.1

Our proof is constructive. We select the columns in a greedy fashion. At every round, we will control the evolution of the eigenvalues. The interior eigenvalues will be controlled by interlacing, for an appropriately chosen induction hypothesis (4.13). The extreme eigenvalues will then be controlled by the secular equation.

Let us now provide some more details on the greedy selection criterion for choosing the next column in the growing extracted submatrix. This criterion will naturally follow from the subsequent analysis. Imagine you start with a matrix Y_r of columns of X and that

$$Y_{r+1} = [Y_r, y_{r+1}] \tag{4.5}$$

where y_{r+1} is the next selected column of X. We can then write

$$Y_{r+1}^t Y_{r+1} \;=\; \left[\begin{array}{c} y_{r+1}^t \\ Y_r^t \end{array} \right] \left[\begin{array}{cc} y_{r+1} & Y_r \end{array} \right] = \left[\begin{array}{cc} 1 & y_{r+1}^t Y_r \\ Y_r^t y_{r+1} & Y_r^t Y_r \end{array} \right], \tag{4.6}$$

and it is well known that the eigenvalues of $Y_{r+1}^t Y_{r+1}$ are the zeros of the secular equation:

$$q(\lambda) := 1 - \lambda + \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{\lambda - \lambda_{k,r}} \;=\; 0. \tag{4.7}$$

Our first goal will be to prove that

$$\sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{\lambda - \lambda_{k,r}} \leq C \; \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} \tag{4.8}$$

for some constant $C > 0$ and for a range of values for λ that contains the largest root of q. This will require a first key ingredient: a non trivial recurrence relationship between all the spectral gaps at the successive stages of the algorithm. Now, replacing (4.7) with

$$g(\lambda) := 1 - \lambda + C \; \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} = 0, \tag{4.9}$$

will provide a control of the increase of the largest eigenvalue of $Y_{r+1}^t Y_{r+1}$ if y_{r+1} is chosen appropriately. One easy idea is to choose y_{r+1} such that

$$y_{r+1} \in \mathrm{argmin}_{x \text{ column of } X} \; \sum_{k=1}^{r} \frac{(v_k^t Y_r^t x)^2}{k}. \tag{4.10}$$

For such a choice, we can guarantee that

$$\sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} \leq \frac{1}{p-r} \sum_{x \text{ remaining column}} \sum_{k=1}^{r} \frac{(v_k^t Y_r^t x)^2}{k} \tag{4.11}$$

and the second key ingredient in our proof will be a tight control of the RHS term in this last inequality (given early in the proof in Lemma 4.1 below).

4.1. The algorithm

Our deterministic sequential column selection procedure can be recast as an algorithm. We now present the detailed structure of this algorithm as follows.

The column selection algorithm

Result: Y_R
Set $\mathcal{V}_0 = \{x_1, \ldots, x_p\}$. Set $r = 1$. Choose $y_1 \in \mathcal{V}_0$ and set $Y_1 = y_1$.
Let $\lambda_{1,1} = \lambda_1(Y_1^t Y_1)$.
while $\lambda_{1,r} \leq 1 + \varepsilon$ *and* $\lambda_{r,r} \geq 1 - \varepsilon$ **do**
  For $k = 1, \ldots, r$, let v_k be a unit eigenvector of $Y_r^t Y_r$ associated with $\lambda_{k,r} := \lambda_k(Y_r^t Y_r)$.
  Set $\mathcal{V}_r := \{x_1, \ldots, x_p\} \setminus \{y_1, \ldots, y_r\}$.
  Choose $y_{r+1} \in \mathcal{V}_r$ so that

$$\sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} \quad \leq \quad \frac{1}{p-r} \sum_{x \in \mathcal{V}_r} \sum_{k=1}^{r} \frac{(v_k^t Y_r^t x)^2}{k}. \tag{4.12}$$

  Set $Y_{r+1} = [Y_r, y_{r+1}]$.
  Set $r = r + 1$.
end

The proof of Theorem 3.1 with ensure that the algorithm above will not stop before having incorporated R columns with R satisfying (3.4).

4.2. First property of y_{r+1}

Lemma 4.1. *For all $r \geq 1$, y_{r+1} verifies*

$$\sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} \quad \leq \quad \frac{\lambda_{1,r} \|X\|^2 \log(r)}{p-r}.$$

Proof. Let X_r be the matrix whose columns are the $x \in \mathcal{V}_r$, i.e. $X_r X_r^t = \sum_{x \in \mathcal{V}_r} x x^t$. Then

$$\sum_{x \in \mathcal{V}_r} (v_k^t Y_r^t x)^2 = \mathrm{Tr}\left(Y_r v_k v_k^t Y_r^t X_r X_r^t\right) \leq \mathrm{Tr}(Y_r v_k v_k^t Y_r^t) \|X_r X_r^t\| \leq \lambda_{k,r} \|X\|^2,$$

which yields the conclusion by plugging in into (4.12) since $\lambda_{k,r} \leq \lambda_{1,r}$. □

4.3. Controlling the individual eigenvalues

Let us define δ as

$$\delta \;=\; \sqrt{\frac{(1+\varepsilon)\|X\|^2 \log R}{p}},$$

so that, from (3.4), $2\delta\sqrt{R} \le \varepsilon$.

Lemma 4.2. *For all r and k with $1 \le k \le r \le R$, we have*

$$1 - \delta\frac{r+k-1}{\sqrt{r}} \;\le\; \lambda_{k,r} \;\le\; 1 + \delta\frac{2r-k}{\sqrt{r}}. \tag{4.13}$$

Proof. It is clear that (4.13) holds for $r = 1$ since then, 1 is the only singular value because the columns are supposed to be normalized.

Assume the induction hypothesis (H_r): for all k with $1 \le k \le r < R$, (4.13) holds.

Let us then show that (H_{r+1}) holds. By the Cauchy interlacing theorem, we have

$$\begin{aligned} \lambda_{k+1,r+1} &\le \lambda_{k,r}, \quad 1 \le k \le r \\ \lambda_{k+1,r+1} &\ge \lambda_{k+1,r}, \quad 0 \le k \le r-1. \end{aligned}$$

Using $(r+1)(2r-k)^2 \le r(2r+1-k)^2$ and $(r+1)(r+k)^2 \le r(r+1+k)^2$, we thus deduce

$$\begin{aligned} \lambda_{k+1,r+1} &\le 1 + \delta\frac{2r-k}{\sqrt{r}} \le 1 + \delta\frac{2(r+1)-(k+1)}{\sqrt{r+1}}, \quad 1 \le k \le r, \\ \lambda_{k+1,r+1} &\ge 1 - \delta\frac{r+k}{\sqrt{r}} \ge 1 - \delta\frac{(r+1)+(k+1)-1}{\sqrt{r+1}}, \quad 0 \le k \le r-1. \end{aligned}$$

It remains to obtain the upper estimate for $\lambda_{1,r+1}$ and the lower one for $\lambda_{r+1,r+1}$. Recall that the eigenvalues of $Y_{r+1}^t Y_{r+1}$ are the zeros of the secular equation:

$$q(\lambda) := 1 - \lambda + \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{\lambda - \lambda_{k,r}} \;=\; 0. \tag{4.14}$$

We first estimate $\lambda_{1,r+1}$ which is the greatest zero of q, and assume for contradiction that

$$\lambda_{1,r+1} > 1 + 2\delta\sqrt{r}. \tag{4.15}$$

From (H_r), we then obtain that for $\lambda \ge 1 + 2\delta\sqrt{r} \ge \lambda_{1,r} + \delta/\sqrt{r}$,

$$q(\lambda) \le 1 - \lambda + \frac{\sqrt{r}}{\delta} \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} := g(\lambda).$$

Let λ^0 be the zero of g. We have $g(\lambda_{1,r+1}) \geq q(\lambda_{1,r+1}) = 0 = g(\lambda^0)$. But g is decreasing, so

$$\lambda_{1,r+1} \leq \lambda^0 = 1 + \frac{\sqrt{r}}{\delta} \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k}.$$

By (H_r), $\lambda_{1,r} \leq 1 + 2\delta\sqrt{R} \leq 1 + \varepsilon$. Thus, using Lemma 4.1 and noting that $r \leq p/2$,

$$\lambda_{1,r+1} \leq 1 + \frac{2\sqrt{r}}{\delta} \frac{(1+\varepsilon)\|X\|^2 \log(R)}{p} = 1 + 2\delta\sqrt{r},$$

which yields a contradiction with the inequality (4.15). Thus, we have that $\lambda_{1,r+1} \leq 1 + 2\delta\sqrt{r}$, and therefore, $\lambda_{1,r+1} \leq 1 + \delta\frac{2r+1}{\sqrt{r+1}}$. This shows that the upper bound in (H_{r+1}) holds.

Finally, to estimate $\lambda_{r+1,r+1}$ which is the smallest zero of q, we write using (H_r) that for $\lambda \leq 1 - 2\delta\sqrt{r} \leq \lambda_{r,r} - \delta/\sqrt{r}$,

$$q(\lambda) \geq 1 - \lambda - \frac{\sqrt{r}}{\delta} \sum_{k=1}^{r} \frac{(v_k^t Y_r^t y_{r+1})^2}{k} := \widetilde{g}(\lambda).$$

By means of the same reasoning as above, we prove by contradiction that $\lambda_{r+1,r+1} \geq 1 - 2\delta\sqrt{r}$, which gives $\lambda_{r+1,r+1} \geq 1 - \delta\frac{2r+1}{\sqrt{r+1}}$ and shows that the lower bound in (H_{r+1}) holds. This completes the proof of Lemma 4.2. □

Applying Lemma 4.2 to $r = R$ and simple algebraic manipulations conclude the proof of Theorem 3.1.

References

[1] Avron, H. and Boutsidis, C., Faster subset selection for matrices and applications. SIAM Journal on Matrix Analysis and Applications, 34 (2013) 4, pp.1464–1499.

[2] Bourgain, J. and Tzafriri, L., Invertibility of "large" submatrices with applications to the geometry of Banach spaces and harmonic analysis. Israel J. Math. 57 (1987), no. 2, 137–224.

[3] Naor, A., Sparse quadratic forms and their geometric applications [following Batson, Spielman and Srivastava]. Séminaire Bourbaki: Vol. 2010/2011. Exposés 1027–1042. Astérisque No. 348 (2012), Exp. No. 1033, viii, 189–217.

[4] Nikolov, A., Kunal T., and Li Z., The geometry of differential privacy: the sparse and approximate cases, Proceedings of the forty-fifth annual ACM symposium on Theory of computing. ACM, (2013

[5] Nikolov, A., and Kunal T., Approximating hereditary discrepancy via small width ellipsoids, In Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms, (2015) 324–336.

[6] Spielman, D. A. and Srivastava, N., An elementary proof of the restricted invertibility theorem. Israel J. Math. 190 (2012), 83–91.

[7] Tropp, J., The random paving property for uniformly bounded matrices, Studia Math., vol. 185, no. 1, pp. 67–82, 2008.

[8] Tropp, J., Norms of random submatrices and sparse approximation. C. R. Acad. Sci. Paris, Ser. I (2008), Vol. 346, pp. 1271-1274.

[9] Vershynin, R., Coordinate restrictions of linear operators in l_2^n. arXiv preprint math/0011232 (2000).

[10] Vershynin, R., John's decompositions: selecting a large part. Israel J. Math. 122 (2001), 253–277.

[11] Youssef, P. A note on column subset selection. Int. Math. Res. Not. IMRN 2014, no. 23, 6431–6447.