Singular value decomposition (SVD) of large random matrices

Marianna Bolla, Katalin Friedl, András Krámli
Budapest University of Technology and Ecomomics
and *University of Szeged*
marib@math.bme.hu

IWFOS'2008
Toulouse, June 19, 2008

Motivation

- New challenge of multivariate statistics: to find linear structures in large real-world data sets like communication, social, cellular networks or microarray measurements.
- To fill the gap between the theory of random matrices and classical multivariate analysis.
- To generalize results of Bolla, Lin. Alg. Appl., 2005 for the SVD of large rectangular random matrices and for the contingency table matrix formed by categorical variables in order to perform two-way clustering of these variables.
- To regard large contingency tables as continuous objects, or to investigate testable parameters of them by randomizing smaller tables out of them.

Motivation

- New challenge of multivariate statistics: to find linear structures in large real-world data sets like communication, social, cellular networks or microarray measurements.
- To fill the gap between the theory of random matrices and classical multivariate analysis.
- To generalize results of Bolla, Lin. Alg. Appl., 2005 for the SVD of large rectangular random matrices and for the contingency table matrix formed by categorical variables in order to perform two-way clustering of these variables.
- To regard large contingency tables as continuous objects, or to investigate testable parameters of them by randomizing smaller tables out of them.

Motivation

- New challenge of multivariate statistics: to find linear structures in large real-world data sets like communication, social, cellular networks or microarray measurements.
- To fill the gap between the theory of random matrices and classical multivariate analysis.
- To generalize results of Bolla, Lin. Alg. Appl., 2005 for the SVD of large rectangular random matrices and for the contingency table matrix formed by categorical variables in order to perform two-way clustering of these variables.
- To regard large contingency tables as continuous objects, or to investigate testable parameters of them by randomizing smaller tables out of them.

Motivation

- New challenge of multivariate statistics: to find linear structures in large real-world data sets like communication, social, cellular networks or microarray measurements.
- To fill the gap between the theory of random matrices and classical multivariate analysis.
- To generalize results of Bolla, Lin. Alg. Appl., 2005 for the SVD of large rectangular random matrices and for the contingency table matrix formed by categorical variables in order to perform two-way clustering of these variables.
- To regard large contingency tables as continuous objects, or to investigate testable parameters of them by randomizing smaller tables out of them.

Notation

Definition

The $m \times n$ real matrix $\mathbf{W}$ is a Wigner-noise if its entries w_{ij} $(1 \leq i \leq m,\ 1 \leq j \leq n)$ are independent random variables, $\mathbb{E}(w_{ij}) = 0$, and the w_{ij}'s are uniformly bounded (i.e., there is a constant $K > 0$, independently of m and n, such that $|w_{ij}| \leq K$, $\forall i, j$).

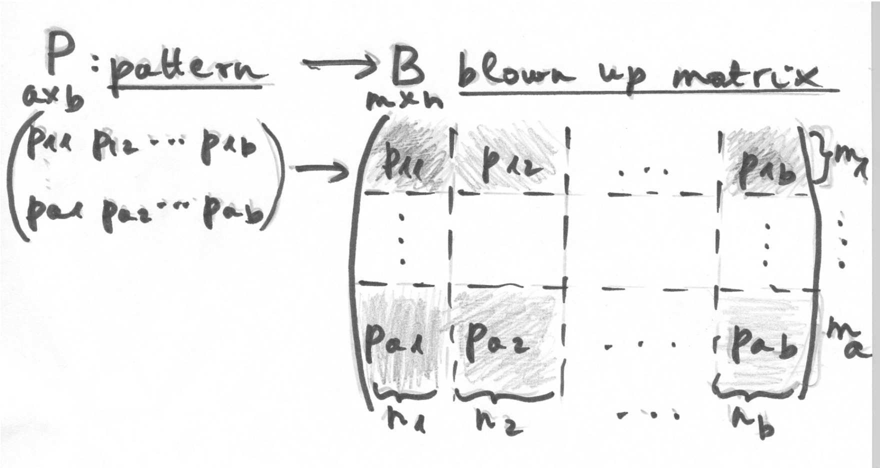
Though, the main results of this paper can be extended to w_{ij}'s with any light-tail distribution (especially to Gaussian distributed w_{ij}'s), our almost sure results will be based on the assumptions of this definition.

Definition

The $m \times n$ real matrix $\mathbf{B}$ is a blown up matrix, if there is an $a \times b$ so-called pattern matrix $\mathbf{P}$ with entries $0 \leq p_{ij} \leq 1$, and there are positive integers $m_1, \dots, m_a$ with $\sum_{i=1}^{a} m_i = m$ and $n_1, \dots, n_b$ with $\sum_{i=1}^{b} n_i = n$, such that the matrix $\mathbf{B}$ can be divided into $a \times b$ blocks, where block (i, j) is an $m_i \times n_j$ matrix with entries equal to p_{ij} $(1 \leq i \leq a,\ 1 \leq j \leq b)$.

Such schemes are sought for in microarray analysis and they are called chess-board patterns, cf. Kluger et al., Genome Research, 2003.

Blown up matrix



The investigated situation

Fix **P**, blow it up to **B**, and **A**:=**B**+**W**.

Almost sure properties of **A** are investigated, when $m_1, \ldots, m_a \to \infty$ and $n_1, \ldots, n_b \to \infty$, roughly speaking, at the same rate.

- Growth Condition 1 There exists a constant $0 < c < 1$ such that $m_i/m \geq c$ $(i = 1, \ldots, a)$ and there exists a constant $0 < d < 1$ such that $n_i/n \geq d$ $(i = 1, \ldots, b)$.
- Growth Condition 2 There exist constants $C \geq 1$, $D \geq 1$, and $C_0 > 0$, $D_0 > 0$ such that $m \leq C_0 \cdot n^C$ and $n \leq D_0 \cdot m^D$ hold for sufficiently large m and n.

The investigated situation

Fix **P**, blow it up to **B**, and **A**:=**B**+**W**.
Almost sure properties of **A** are investigated, when $m_1, \ldots, m_a \to \infty$ and $n_1, \ldots, n_b \to \infty$, roughly speaking, at the same rate.

- Growth Condition 1 There exists a constant $0 < c < 1$ such that $m_i/m \geq c$ $(i = 1, \ldots, a)$ and there exists a constant $0 < d < 1$ such that $n_i/n \geq d$ $(i = 1, \ldots, b)$.
- Growth Condition 2 There exist constants $C \geq 1$, $D \geq 1$, and $C_0 > 0$, $D_0 > 0$ such that $m \leq C_0 \cdot n^C$ and $n \leq D_0 \cdot m^D$ hold for sufficiently large m and n.

Almost sure properties of SVD

Definition

Property $\mathcal{P}_{m,n}$ holds for $\mathbf{A}_{m \times n}$ almost surely (with probability 1) if $\mathbb{P}\left(\exists\ m_0, n_0 \in \mathbb{N}\ \text{such that for}\ m \geq m_0\ n \geq n_0\ \mathbf{A}_{m \times n}\ \text{has}\ \mathcal{P}_{m,n}\right) = 1$. Here we may assume GC1 or GC2 for the growth of m and n, while K is kept fixed.

Füredi, Komlós, Combinatorica, 1981 $\longrightarrow$ Achlioptas, McSherry, Proc. ACM, 2001 $\longrightarrow$ $\|\mathbf{W}\| = \mathcal{O}(\sqrt{m+n})$ in probability.
N. Alon et al., Israel J. Math., 2002 + Borel–Cantelli Lemma $\longrightarrow$

Lemma

There exist positive constants C_{K1} and C_{K2}, depending on the common bound on the entries of $\mathbf{W}$, such that

$$\mathbb{P}\left(\|\mathbf{W}\| > C_{K1} \cdot \sqrt{m+n}\right) \leq \exp[-C_{K2} \cdot (m+n)].$$

Alon's sharp concentration theorem

Theorem

$\widetilde{\mathbf{W}}$ is $q \times q$ real symmetric matrix, its entries in and above the main diagonal are independent random variables with absolute value at most 1. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$: eigenvalues of $\widetilde{\mathbf{W}}$. For any $t > 0$:

$$\mathbb{P}\left(|\lambda_i - \mathbb{E}(\lambda_i)| > t\right) \leq \exp\left(-\frac{(1-o(1))t^2}{32i^2}\right) \quad \textit{when} \quad i \leq \frac{q}{2},$$

and the same estimate holds for the probability

$$\mathbb{P}\left(|\lambda_{q-i+1} - \mathbb{E}(\lambda_{q-i+1})| > t\right).$$

generalization for rectangular matrices

$\mathbf{W}$ Wigner-noise, $|w_{ij}| \leq K$, $\forall i, j$.

$$\widetilde{\mathbf{W}} = \frac{1}{K} \cdot \begin{pmatrix} \mathbf{0} & W \\ W^T & \mathbf{0} \end{pmatrix}$$

satisfies the conditions of the theorem, its largest and smallest eigenvalues:

$$\lambda_i(\widetilde{\mathbf{W}}) = -\lambda_{n+m-i+1}(\widetilde{\mathbf{W}}) = \frac{1}{K} \cdot s_i(\mathbf{W}), \qquad i = 1, \dots, \min\{m, n\},$$

the others are zeros.

Singular values of a noisy matrix

Under the usual growth condition, all the $r = \operatorname{rank} \mathbf{P} \le \min\{a, b\}$ non-zero singular values of the $m \times n$ blown-up matrix $\mathbf{B}$ are of order $\sqrt{mn}$.

Theorem

Let $\mathbf{A} = \mathbf{B} + \mathbf{W}$ be an $m \times n$ random matrix, where $\mathbf{B}$ is a blown up matrix with positive singular values $s_1, \dots, s_r$ and $\mathbf{W}$ is a Wigner-noise of the same size. Then the matrix $\mathbf{A}$ almost surely has r singular values $z_1, \dots, z_r$ with $|z_i - s_i| = \mathcal{O}(\sqrt{m+n})$, $i = 1, \dots, r$, and for the other singular values $z_j = \mathcal{O}(\sqrt{m+n})$, $j = r+1, \dots, \min\{m, n\}$ hold almost surely, as $m, n \to \infty$ under GC1.

Classification via singular vector pairs

$\mathbf{Y} := (\mathbf{y}_1, \dots, \mathbf{y}_r)$ $m \times r$ left singular vectors of $\mathbf{A}$.
Rows of $\mathbf{Y}$: $\mathbf{y}^1, \dots, \mathbf{y}^m \in \mathbb{R}^r \rightarrow$ genes' representatives.
$\mathbf{X} := (\mathbf{x}_1, \dots, \mathbf{x}_r)$ $n \times r$ right singular vectors of $\mathbf{A}$.
Rows of $\mathbf{X}$: $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathbb{R}^r \rightarrow$ conditions' representatives.

$$S_a^2(\mathbf{Y}) := \sum_{i=1}^{a} \sum_{j \in A_i} \|\mathbf{y}^j - \bar{\mathbf{y}}^i\|^2, \quad \text{where} \quad \bar{\mathbf{y}}^i = \frac{1}{m_i} \sum_{j \in A_i} \mathbf{y}^j,$$

$$S_b^2(\mathbf{X}) := \sum_{i=1}^{b} \sum_{j \in B_i} \|\mathbf{x}^j - \bar{\mathbf{x}}^i\|^2, \quad \text{where} \quad \bar{\mathbf{x}}^i = \frac{1}{n_i} \sum_{j \in B_i} \mathbf{x}^j.$$

Theorem

$$S_a^2(\mathbf{Y}) = \mathcal{O}\left(\frac{m+n}{mn}\right) \quad \textit{and} \quad S_b^2(\mathbf{X}) = \mathcal{O}\left(\frac{m+n}{mn}\right)$$

almost surely, for the a- and b-variances of the representatives.

Perturbation results for correspondence matrices

$\mathbf{P}$: $a \times b$ contingency table (nonnegative, uniformly bounded entries). $\mathbf{B}$:$m \times n$ blown up contingency table.
Correspondence analysis: to find maximally correlated factors with respect to the marginal distributions of the two underlying categorical variables. Benzécri et al., Dunod, Paris, 1973.
The categories may be measured in different units $\longrightarrow$ normalization: correspondence transformation $\longrightarrow$ $\mathbf{B}_{corr}$ has entries in [0,1] and maximum singular value 1.
Proposition: Under GC1 and GC2, there is a significant gap between the r largest (where $k = \text{rank}\,(\mathbf{B}) = \text{rank}\,(\mathbf{P})$) and the other singular values of $\mathbf{A}_{corr}$, the matrix obtained from the noisy matrix $\mathbf{A}$=$\mathbf{B}$+$\mathbf{W}$ by the correspondence transformation.

$$\mathbf{B}_{corr} := \mathbf{D}_{Brow}^{-1/2}\mathbf{B}\mathbf{D}_{Bcol}^{-1/2} \quad \text{and} \quad \mathbf{A}_{corr} := \mathbf{D}_{Arow}^{-1/2}\mathbf{A}\mathbf{D}_{Acol}^{-1/2}$$

Noisy correspondence vector pairs

$$\mathbf{y}_{corr\ i} := \mathbf{D}_{Arow}^{-1/2}\mathbf{y}_i, \qquad \mathbf{x}_{corr\ i} := \mathbf{D}_{Acol}^{-1/2}\mathbf{x}_i \quad (i = 1, \dots, r).$$

a- and b-variances of the representatives:

$$S_a^2(\mathbf{Y}_{corr}) = \sum_{i=1}^{a} \sum_{j \in A_i} d_{Arow\ j} \|\mathbf{y}_{corr}^j - \bar{\mathbf{y}}_{corr}^i\|^2, \quad \bar{\mathbf{y}}_{corr}^i = \sum_{j \in A_i} d_{Arow\ j}\mathbf{y}_{corr}^j$$

$$S_b^2(\mathbf{X}_{corr}) = \sum_{i=1}^{b} \sum_{j \in B_i} d_{Acol\ j} \|\mathbf{x}_{corr}^j - \bar{\mathbf{x}}_{corr}^i\|^2, \quad \bar{\mathbf{x}}_{corr}^i = \sum_{j \in B_i'} d_{Acol\ j}\mathbf{x}_{corr}^j$$

$$S_a^2(\mathbf{Y}_{corr}), \quad S_b^2(\mathbf{X}_{corr}) = \mathcal{O}(\max\{n^{-\tau}, m^{-\tau}\} \quad 0 < \tau < 1$$

Recognizing the structure

Theorem

Let $\mathbf{A}_{m\times n}$ be a sequence of $m \times n$ matrices, where m and n tend to infinity. Assume, that $\mathbf{A}_{m\times n}$ has exactly k singular values of order greater than $\sqrt{m+n}$ (k is fixed). If there are integers $a \geq k$ and $b \geq k$ such that the a- and b-variances of the row- and column-representatives are $\mathcal{O}(\frac{m+n}{mn})$, then there is a blown up matrix $\mathbf{B}_{m\times n}$ such that $\mathbf{A}_{m\times n} = \mathbf{B}_{m\times n} + \mathbf{E}_{m\times n}$, with $\|\mathbf{E}_{m\times n}\| = \mathcal{O}(\sqrt{m+n})$.

The proof gives an explicit construction for $\mathbf{B}_{m\times n}$ by means of metric classification methods. For SVD of large rectangular matrices: randomized algorithms, e.g., A. Frieze and R. Kannan, Combinatorica, 1999.

Szemerédi's Lemma for rectangular arrays

Lemma

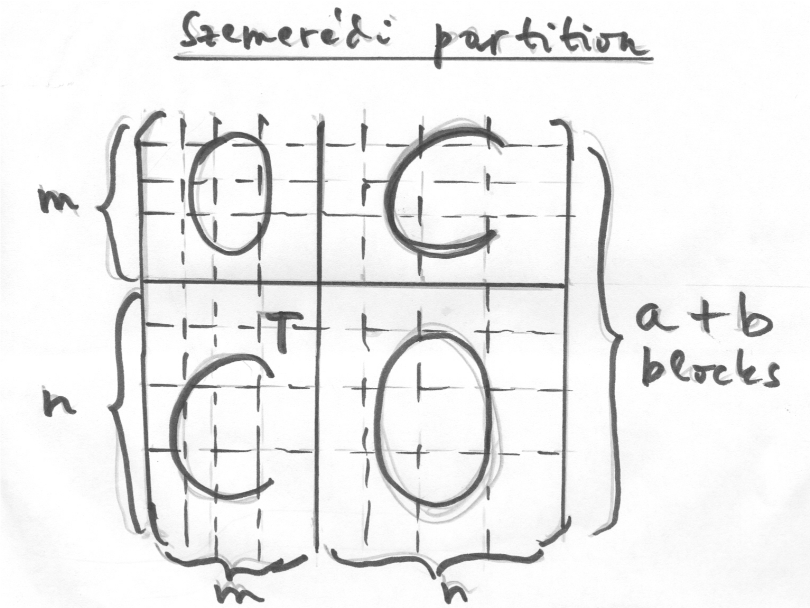
$\forall \varepsilon > 0$ and $\mathbf{C}_{m\times n}$ $\exists \mathbf{B}_{m\times n}$ blown up matrix of pattern matrix $\mathbf{P}_{a\times b}$ with $a + b \le 4^{1/\varepsilon^2}$ (independently of m, n) such that

$$\|\mathbf{C} - \mathbf{B}\|_\square \le \varepsilon \|\mathbf{C}\|_2.$$

Here $\|\mathbf{C} - \mathbf{B}\|_\square = \max_{A\subset\{1,\dots,m\},\, B\subset\{1,\dots,n\}} \frac{1}{mn} \sum_{i\in A} \sum_{j\in B} |c_{ij} - b_{ij}|$ and $\|\mathbf{C}\|_2 = \sqrt{\frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}^2}$.
Proof: apply the Lovász's version of the lemma to

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{0} \end{pmatrix} \; (m+n)\times(m+n) \text{ weight matrix of a weighted graph.}$$

Szemerédi partition of a rectangular array



Convergence of contingency tables

$\mathbf{C}_{m \times n}$: contingency table, $0 \leq c_{ij} \leq 1$
$\mathbf{F}_{a \times b}$: fixed "small" 0/1 table.
Randomize an $a \times b$ table of 0/1's out of $\mathbf{C}$: choose a rows and b columns randomly, then choose the entries conditionally independently with $\mathbb{P}(1) = c_{ij}$, $\mathbb{P}(0) = 1 - c_{ij}$ in the ij-th position.
It can be reached with adding an appropriate Wigner-noise.

$$\mathbb{P}(\text{randomized table} = \mathbf{F}) = \sum_{\Phi,\Psi} \frac{1}{m^a n^b} \prod_{f_{ij}=1} c_{\Phi(i),\Psi(j)} \prod_{f_{ij}=0} (1 - c_{\Phi(i),\Psi(j)}),$$

$$\mathbf{F} \to \mathbf{C} \quad \text{homomorphism's} \quad \mathrm{dens}\,(\mathbf{F}, \mathbf{C}) := \sum_{\Phi,\Psi} \frac{1}{m^a n^b} \prod_{f_{ij}=1} c_{\Phi(i),\Psi(j)},$$

where $\Phi : Row_F \to Row_C$, $\Psi : Col_F \to Col_C$ are injective maps.

Definition

$\mathbf{C}_{m,n}$ is convergent, if $\mathrm{dens}\,(\mathbf{F}, \mathbf{C}_{m,n})$ converges, $\forall \mathbf{F}$.

Remark: $\mathbf{C}_{m,n}$'s are more and more similar in small details.

Testable contingency table parameters

Limit object: contingon (non-negative, bounded function on $[0,1] \times [0,1]$), generalization of graphons, cf. L. Lovász and B. Szegedy, J. Combin. Theory, 2006.

Contingon, belonging to $\mathbf{C}_{m\times n}$: stepwise constant function. If $m, n \to \infty$, it becomes a continuous object.

Definition

The contingency table parameter f is testable if $f(\mathbf{C}_{m,n})$ converges, whenever $\mathbf{C}_{m,n}$ converges.

Remark: f reflects some statistical property, invariant under isomorphism of the contingency table and scale of the entries.

Conclusion: to find a good approximation of $f(\mathbf{C}_{m\times n})$ with m and n "large", it is enough to appropriately randomize a "smaller" contingency table out of $\mathbf{C}$.