Statistical Inference on Large Contingency Tables: Convergence, Testability, Stability

Marianna Bolla
Institute of Mathematics
Budapest University of Technology and Economics
marib@math.bme.hu

India, December, 2010

Motivation

- To recover the structure of large rectangular arrays, for example, microarrays, socal, economic, or communication networks, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes.
  Two directions:
- 1. Select a smaller part (by an appropriate randomization) and process SVD or correspondence analysis on it.
- 2. Regard it as a continuous object and set up a bilinear programming task with constraints. In this way, fuzzy clusters are obtained.

Motivation

- To recover the structure of large rectangular arrays, for example, microarrays, socal, economic, or communication networks, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes.
  Two directions:
- 1. Select a smaller part (by an appropriate randomization) and process SVD or correspondence analysis on it.
- 2. Regard it as a continuous object and set up a bilinear programming task with constraints. In this way, fuzzy clusters are obtained.

Motivation

- To recover the structure of large rectangular arrays, for example, microarrays, socal, economic, or communication networks, classical methods of cluster and correspondence analysis may not be carried out on the whole table because of computational size limitations. In other situations, we want to compare contingency tables of different sizes.
  Two directions:
- 1. Select a smaller part (by an appropriate randomization) and process SVD or correspondence analysis on it.
- 2. Regard it as a continuous object and set up a bilinear programming task with constraints. In this way, fuzzy clusters are obtained.

References

- We generalize some theorems of Borgs, Chayes, Lovász, Sós, Vesztergombi, Convergent graph sequences I: subgraph sequences, metric properties and testing, Advances in Math. 2008 to rectangular arrays and to testable parameters defined on them.
- In Bolla, Friedl, Krámli, Singular value decomposition of large random matrices (for two-way classification of microarrays), Journal of Multivariate Analysis 101, 2010 we investigated effects of random perturbations on the entries to the singular spectrum, clustering effect, and correspondence factors.

References

- We generalize some theorems of Borgs, Chayes, Lovász, Sós, Vesztergombi, Convergent graph sequences I: subgraph sequences, metric properties and testing, Advances in Math. 2008 to rectangular arrays and to testable parameters defined on them.
- In Bolla, Friedl, Krámli, Singular value decomposition of large random matrices (for two-way classification of microarrays), Journal of Multivariate Analysis 101, 2010 we investigated effects of random perturbations on the entries to the singular spectrum, clustering effect, and correspondence factors.

Notation

Let $C = C_{m \times n}$ be a contingency table of row set $Row_C = \{1, \ldots, m\}$ and column set $Col_C = \{1, \ldots, n\}$.
c_{ij}'s are interactions between the rows and columns, and they are normalized such that $0 \leq c_{ij} \leq 1$.
Binary table: $0/1$ entries.
Row-weights: $\alpha_1, \ldots, \alpha_m \geq 0$
Column-weights: $\beta_1, \ldots, \beta_n \geq 0$
(Individual importance of the categories. In correspondence analysis, these are the marginals.)

A contingency table is called simple if all the row- and column-weights are equal to 1.
Assume that C does not contain identically zero rows or columns, moreover C is dense in the sense that the number of nonzero entries is comparable with mn. Let $\mathcal{C}$ denote the set of such tables (with any natural numbers m and n).
Consider a simple binary table $F_{a\times b}$ and maps $\Phi : Row_F \to Row_C$, $\Psi : Col_F \to Col_C$; further

$$\alpha_\Phi := \prod_{i=1}^{a} \alpha_{\Phi(i)}, \quad \beta_\Psi := \prod_{j=1}^{b} \beta_{\Psi(j)}, \quad \alpha_C := \sum_{i=1}^{m} \alpha_i, \quad \beta_C := \sum_{j=1}^{n} \beta_j.$$

Homomorphism density

Definition

The $F \to C$ homomorphism density is

$$t(F,C) = \frac{1}{(\alpha_C)^a(\beta_C)^b} \sum_{\Phi,\Psi} \alpha_\Phi \beta_\Psi \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}.$$

If C is simple, then

$$t(F,C) = \frac{1}{m^a n^b} \sum_{\Phi,\Psi} \prod_{f_{ij}=1} c_{\Phi(i)\Psi(j)}.$$

In addition, if C is binary too, then $t(F,C)$ is the probability that a random map $F \to C$ is a homomorphism (preserves the 1's).

The maps Φ and Ψ correspond to sampling a rows and b columns out of Row_C and Col_C with replacement, respectively. In case of simple C it means uniform sampling, otherwise the rows and columns are selected with probabilities proportional to their weights.

The following simple binary random table $\xi(a \times b, C)$ will play an important role in proving the equivalent theorems of testability.

Select a rows and b columns of C with replacement, with probabilities α_i/α_C $(i = 1, \dots, m)$ and β_j/β_C $(j = 1, \dots, n)$, respectively. If the ith row and jth column of C are selected, they will be connected by 1 with probability c_{ij} and 0, otherwise, independently of the other selected row–column pairs, conditioned on the selection of the rows and columns.

For large m and n, $\mathbb{P}(\xi(a \times b, C) = F)$ and $t(F, C)$ are close to each other.

Definition

Definition

We say that the sequence $(C_{m\times n})$ of contingency tables is convergent if the sequence $t(F, C_{m\times n})$ converges for any simple binary table F as $m, n \to \infty$.

The convergence means that the tables $C_{m\times n}$ become more and more similar in small details as they are probed by smaller 0-1 tables $(m, n \to \infty)$.

The limit object

The limit object is a measurable function $U : [0,1]^2 \to [0,1]$ and we call it contingon.

In the $m = n$ and symmetric case, C can be regarded as the weight matrix of an edge- and node-weighted graph (the row-weights are equal to the column-weights, loops are possible) and the limit object was introduced as graphon, see Borgs et al.

The step-function contingon U_C is assigned to C in the following way: the sides of the unit square are divided into intervals $I_1, \dots, I_m$ and $J_1, \dots, J_n$ of lengths $\alpha_1/\alpha_C, \dots, \alpha_m/\alpha_C$ and $\beta_1/\beta_C, \dots, \beta_n/\beta_C$, respectively; then over the rectangle $I_i \times J_j$ the step-function takes on the value c_{ij}.

The metric inducing the convergence

Definition

The cut distance between the contingons U and V is

$$\delta_{\square}(U,V) = \inf_{\mu,\nu} \|U - V^{\mu,\nu}\|_{\square}$$

where the cut norm of the contingon U is defined by

$$\|U\|_{\square} = \sup_{S,T\subset[0,1]} \left| \iint_{S\times T} U(x,y)\,dx\,dy \right| ,$$

and the infimum is taken over all measure preserving bijections $\mu,\nu : [0,1] \to [0,1]$, while $V^{\mu,\nu}$ denotes the transformed V after performing the measure preserving bijections μ and ν on the sides of the unit square, respectively.

Equivalence classes of contingons

An equivalence relation is defined over the set of contingons: two contingons belong to the same class if they can be transformed into each other by measure preserving map, i.e., their cut distance is zero.

In the sequel, we consider contingons modulo measure preserving maps, and under contingon we understand the whole equivalence class. By a theorem of Borgs et al. (2008), the equivalence classes form a compact metric space with the $\delta_\square$ metric.

Distance of contingency tables of different sizes

Definition

The cut distance between the contingency tables $C, C' \in \mathcal{C}$ is

$$\delta_\square(C, C') = \delta_\square(U_C, U_{C'}).$$

By the above remarks, the distance of C and C' is indifferent to permutations of the rows or columns of C and C'. In the special case when C and C' are of the same size, $\delta_\square(C, C')$ is $\frac{1}{mn}$ times the usual cut distance of matrices, cf. Frieze and Kannan (1999).

Uniqueness of the limit

The following reversible relation between convergent contingency table sequences and contingons also holds, as a rectangular analogue of a theorem of Borgs et al. (2008).

Theorem

For any convergent sequence $(C_{m\times n}) \subset \mathcal{C}$ there exists a contingon such that $\delta_\square(U_{C_{m\times n}}, U) \to 0$ as $m, n \to \infty$. Conversely, any contingon can be obtained as the limit of a sequence of contingency tables in $\mathcal{C}$. The limit of a convergent contingency table sequence is essentially unique: if $C_{m\times n} \to U$, then also $C_{m\times n} \to U'$ for precisely those contingons U' for which $\delta_\square(U, U') = 0$.

It also follows that a sequence of contingency tables in $\mathcal{C}$ is convergent if, and only if it is a Cauchy sequence in the metric $\delta_\square$.

Randomization

A simple binary random $a \times b$ table $\xi(a \times b, U)$ can also be randomized based on the contingon U in the following way. Let $X_1, \dots, X_a$ and $Y_1, \dots, Y_b$ be i.i.d., uniformly distributed random numbers on [0,1]. The entries of $\xi(a \times b, U)$ are indepenent Bernoully random variables, namely the entry in the ith row and jth column is 1 with probability $U(X_i, Y_j)$ and 0, otherwise.
It is easy to see that the distribution of the previously defined $\xi(a \times b, C)$ and that of $\xi(a \times b, U_C)$ is the same.
It is important that

$$\mathbb{P}\left(\delta_\square(U, \xi(a \times b, U)) < \frac{10}{\sqrt{\log_2(a+b)}}\right) \geq 1 - e^{-\frac{(a+b)^2}{2\log_2(a+b)}}$$

that is true for $U_{C_{m \times n}}$ independently of m, n.

Exchangeable random arrays

Note, that in the above way, we can as well randomize an infinite simple binary table $\xi(\infty \times \infty, U)$ out of the contingon U by generating countably infinitely many i.i.d. uniform random numbers on [0,1]. The distribution of the infinite binary array $\xi(\infty \times \infty, U)$ is denoted by $\mathbb{P}_U$.

Because of the symmetry of the construction, this is an exchangeable array in the sense that the joint distribution of its entries is invariant under permutations of the rows and colums.

Moreover, any exchangeable binary array is a mixture of such $\mathbb{P}_U$'s.

More precisely, the Aldous–Hoover (Kallenberg) Representation Theorem (Representations for partially exchangeable arrays of random variables, J. Multivar. Anal. 1981) states that for every infinite exchangeable binary array ξ there is a probability distribution μ (over the contingons) such that $\mathbb{P}(\xi \in A) = \int \mathbb{P}_U(A)\, \mu(dU)$.

Definition of testability

A function $f : C \to \mathbb{R}$ is called a contingency table parameter if it is invariant under isomorphism and scaling of the rows/columns. In fact, it is a statistic evaluated on the table, and hence, we are interested in contingency table parameters that are not sensitive to minor changes in the entries of the table.

Definition

A contingency table parameter f is testable if for every $\varepsilon > 0$ there are positive integers a and b such that if the row- and column-weights of C satisfy

$$\max_i \frac{\alpha_i}{\alpha_C} \le \frac{1}{a}, \qquad \max_j \frac{\beta_j}{\beta_C} \le \frac{1}{b},$$

$$\text{then} \quad \mathbb{P}(|f(C) - f(\xi(a \times b, C))| > \varepsilon) \le \varepsilon.$$

Such a contingency table parameter can be consistently estimated based on a fairly large sample.

Equivalent statements of testability

Theorem

For a testable c. t. parameter f the following are equivalent:

- *For every $\varepsilon > 0$ there are positive integers a and b such that for every contingency table $C \in \mathcal{C}$ with no dominant row- and column-weights,*

$$|f(C) - \mathbb{E}(f(\xi(a \times b, C)))| \leq \varepsilon.$$

- *For every convergent sequence $(C_{m \times n})$ of contingency tables with no dominant row- or columnn-weights, $f(C_{m \times n})$ is also convergent $(m, n \to \infty)$.*
- *f can be extended to contingons such that the extended functional $\tilde{f}$ is continuous in the cut-norm and $\tilde{f}(U_{C_{m \times n}}) - f(C_{m \times n}) \to 0$, whenever $\max_i \alpha_i / \alpha_C \to 0$ and $\max_j \alpha_j / \alpha_C \to 0$ as $m, n \to \infty$.*
- *f is continuous in the cut metric.*

Examples

For example, in case of simple binary tables the singular spectrum is testable, as $C_{m\times n}$ can be regarded as part of the adjacency matrix of a bipartite graph on $m+n$ vertices, where Row_C and Col_C are the two independent vertex sets; further, the ith vertex of Row_C and the jth vertex of Col_C are connected by an edge if and only if $c_{ij}=1$. The non-zero real eigenvalues of the symmetric $(m+n)\times(m+n)$ adjacency matrix of this bipartite graph are the numbers $\pm s_1,\dots,\pm s_r$, where $s_1,\dots,s_r$ are the non-zero singular values of C, and $r\le\min\{m,m\}$ is the rank of C. Consequently, the convergence of adjacency spectra implies the convergence of the singular spectra.

By the Equivalence Theorem, any property of a large contingency table based on its singular value decomposition (e.g., correspondence decomposition) can be concluded from a smaller part of it. In the last section, testability of some balanced classification properties is discussed.

Noisy contingency tables

Definition

The $m \times n$ random matrix E is a noise matrix if its entries are independent, uniformly bounded random variables of zero expectation.

Theorem

The cut norm of any sequence $(E_{m\times n})$ of noise matrices tends to zero as $m, n \to \infty$, almost surely.

Definition

The $m \times n$ real matrix B is a blown up matrix, if there is an $a \times b$ so-called *pattern matrix* P with entries $0 \le p_{ij} \le 1$, 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 B, after rearranging its rows and columns, can be divided into $a \times b$ blocks, where block (i, j) is an $m_i \times n_j$ matrix with entries all equal to p_{ij}.

Let us fix the matrix $P_{a \times b}$, blow it up to obtain matrix $B_{m \times n}$, and let $A_{m \times n} = B + E$, where $E_{m \times n}$ is a noise matrix. If the block sizes grow proportionally, the following almost sure statements are proved in Bolla et. al (2010): the noisy matrix A has as many structural (outstanding) singular values of order $\sqrt{mn}$ as the rank of the pattern matrix, all the other singular values are of order $\sqrt{m+n}$; further, by representing the rows and columns by means of the singular vector pairs corresponding to the structural singular values, the a- and b-variances of the representatives tend to 0 as $m, n \to \infty$.

Convergence of noisy tables

Theorem

Let the block sizes of the blown up matrix $B_{m\times n}$ are $m_1,\dots,m_a$ horizontally, and $n_1,\dots,n_b$ vertically ($\sum_{i=1}^{a} m_i = m$ and $\sum_{j=1}^{b} n_j = n$). Let $A_{m\times n} := B + E$ and $m,n\to\infty$ is such a way that $m_i/m \to r_i$ ($i=1,\dots,a$), $n_j/n\to q_j$ ($j=1,\dots,b$), where r_i's and q_j's are fixed ratios. Under these conditions, the "noisy" sequence $(A_{m\times n})$ converges almost surely.

Conversely, in the presence of structural singular values, with some additional conditions for the representatives, the block structure can be recovered.

Homogeneous partitions

In many applications we are looking for clusters of the rows and columns of a rectangular array such that the densities within the cross-products of the clusters be homogeneous. E.g., in microarray analysis we are looking for clusters of genes and conditions such that genes of the same cluster equally influence conditions of the same cluster. The following theorem ensures the existence of such a structure with possibly many clusters. However, the number of clusters does not depend on the size of the array, it merely depends on the accuracy of the approximation.

Theorem

For every $\varepsilon > 0$ and $C_{m\times n} \in \mathcal{C}$ there exists a blown up matrix $B_{m\times n}$ of an $a \times b$ pattern matrix with $a + b \le 4^{1/\varepsilon^2}$ (independently of m and n) such that $\delta_{\square}(C, B) \le \varepsilon$.

The theorem is a consequence of the Szemerédi's Regularity Lemma (see Frieze and Kannan (1999), Borgs et al. (2008)) and can be proved by embedding C into the adjacency matrix of an edge-weighted bipartite graph. The statement of the theorem is closely related to the testability of the following contingency table parameter:

$$S^2_{a,b}(C) = \min \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k \in A_i} \sum_{l \in B_j} (c_{kl} - \bar{c}_{ij})^2, \quad \bar{c}_{ij} = \frac{1}{|A_i| \cdot |B_j|} \sum_{k \in A_i} \sum_{l \in B_j} c_{kl}$$

where the minimum is taken over balanced a- and b-partitions $A_1, \dots, A_a$ and $B_1, \dots, B_b$ of Row_C and Col_C, respectively; further, instead of c_{kl} we may take $\alpha_k \beta_l c_{kl}$ in the row- and column-weighted case, provided there are no dominant rows/columns.

Partitions of contingons

As $S^2_{a,b}(C)$ is a testable contingency table parameter, by the Equivalence Theorem, it can be continuously extended to contingons:

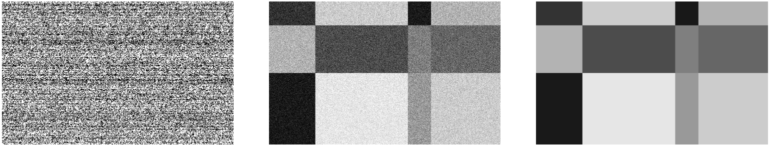
$$S^2_{a,b}(U) = \min \sum_{i=1}^{a} \sum_{j=1}^{b} \int_{A_i \times B_j} (U(x,y) - \bar{U}_{ij})^2 \, dxdy, \; \bar{U}_{ij} = \frac{\int_{A_i \times B_j} U(x,y) \, dxdy}{\lambda(A_i) \cdot \lambda(B_j)}$$

and the minimum is taken over balanced a- and b-partitions $A_1, \dots, A_a$ and $B_1, \dots, B_b$ of the $[0,1]$ interval into measurable subsets, respectively (λ is the Lebesgue measure).
Minimizing $S^2_{a,b}(U_C)$ is a bilinear programming task in the variables $x_{ij} = \lambda(A_i \cap I_j)$ ($i = 1, \dots, a$; $j = 1, \dots, m$) and $y_{ij} = \lambda(B_i \cap J_j)$ ($i = 1, \dots, b$; $j = 1, \dots, n$) under constraints of balance.
As for large m, n $S^2_{a,b}(U_C)$ is very close to $S^2_{a,b}(C)$, the solution of the continuous problem gives fuzzy clusters.

Application

We applied our spectral partitioning algorithm for mixture of noisy data: $a = 3$, $b = 4$, $m_1 = 3, m_2 = 2, m_3 = 1$, $n_1 = 2, n_2 = 4, n_3 = 1, n_4 = 3$. After the starting blow up: 6×10 table, then its $5, 10, \dots, 100$-fold blown up tables with noise are presented.

- the 300×500 noisy table
- the 600×1000 blown up table, with rows and columns sorted according to their cluster memberships obtained by k-means algorithm
- the colour illustration of the average densities of the blocks formed by low rank approximation via SVD

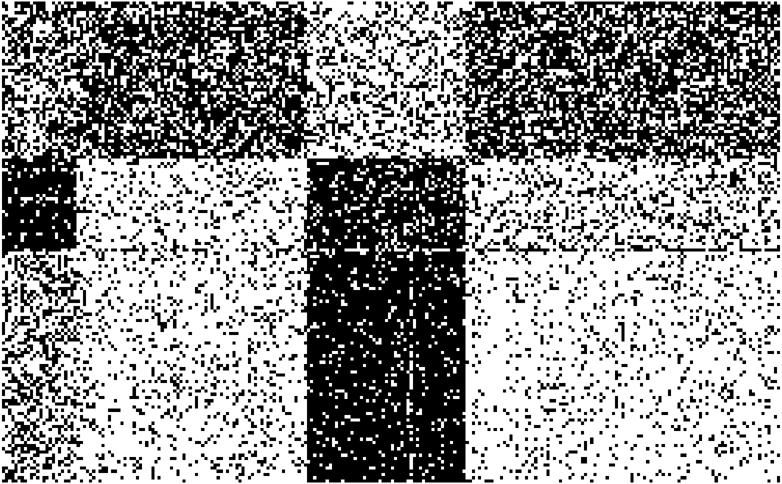


5-fold blow up

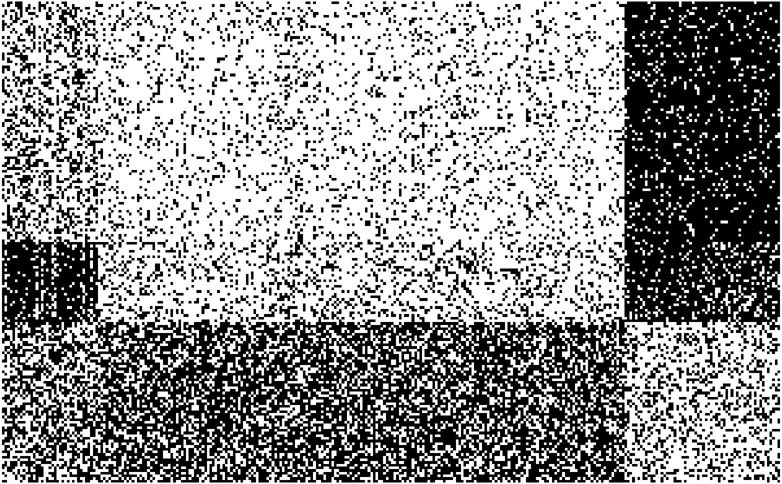
10-fold blow up

15-fold blow up

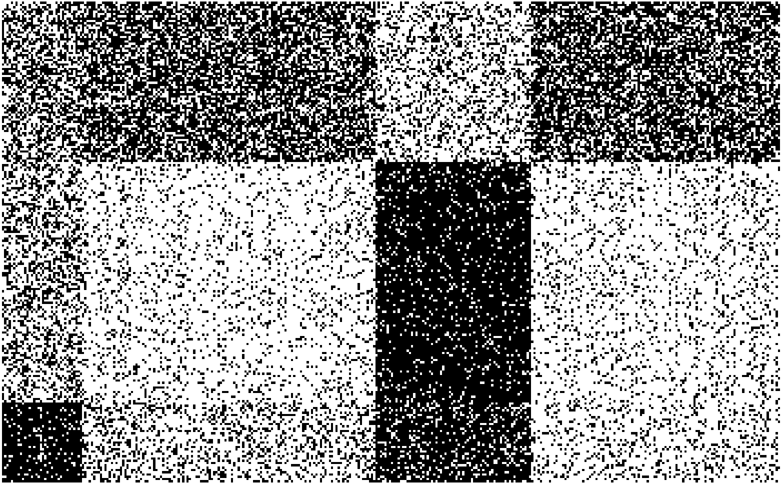
20-fold blow up



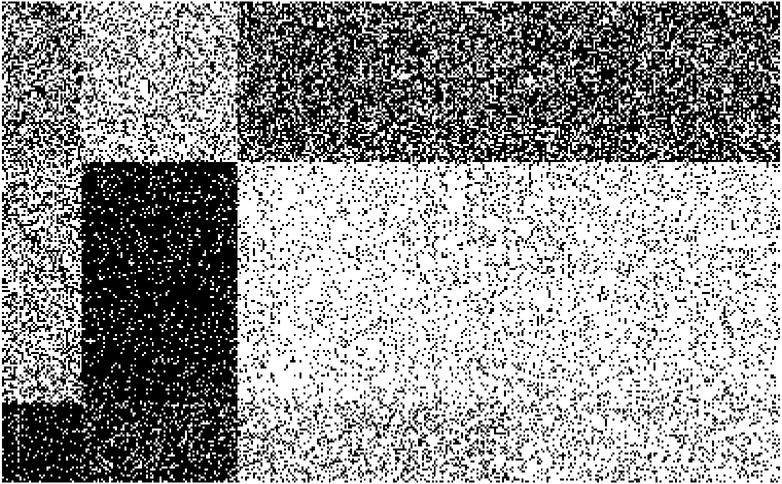
25-fold blow up



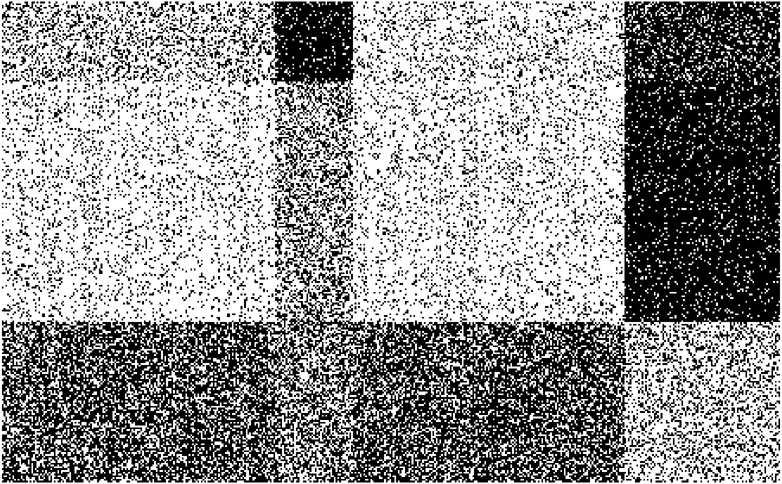
30-fold blow up



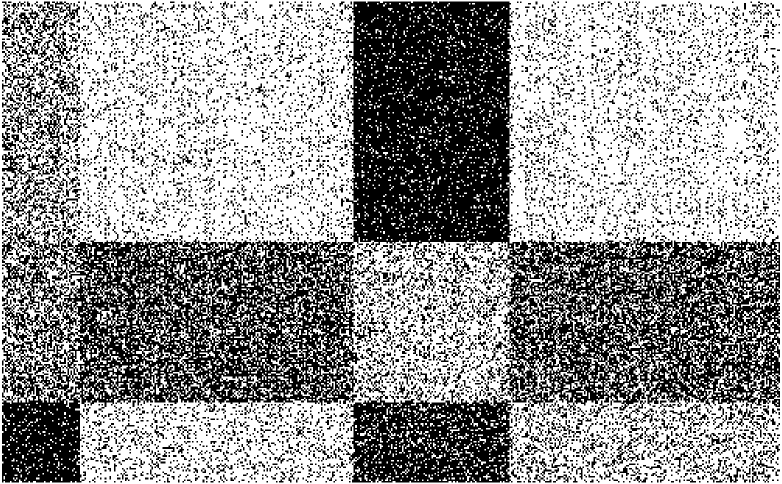
35-fold blow up



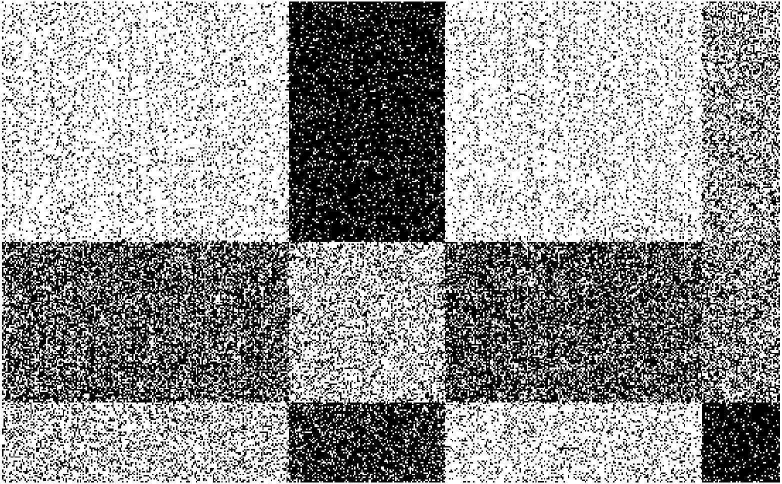
40-fold blow up



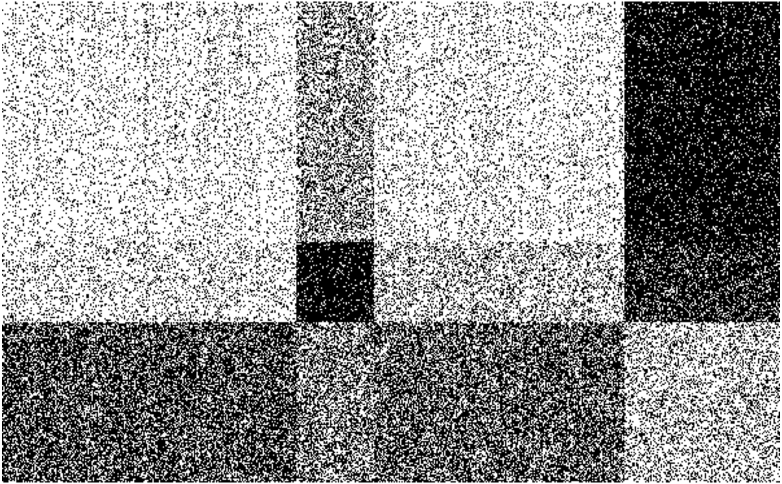
45-fold blow up



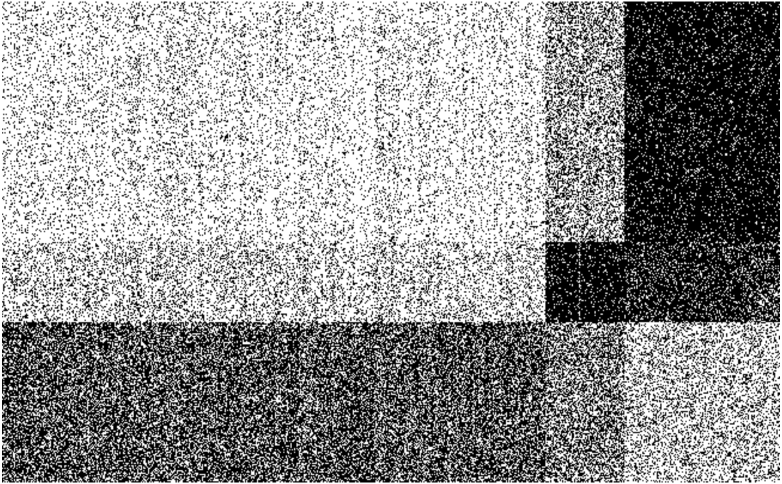
50-fold blow up



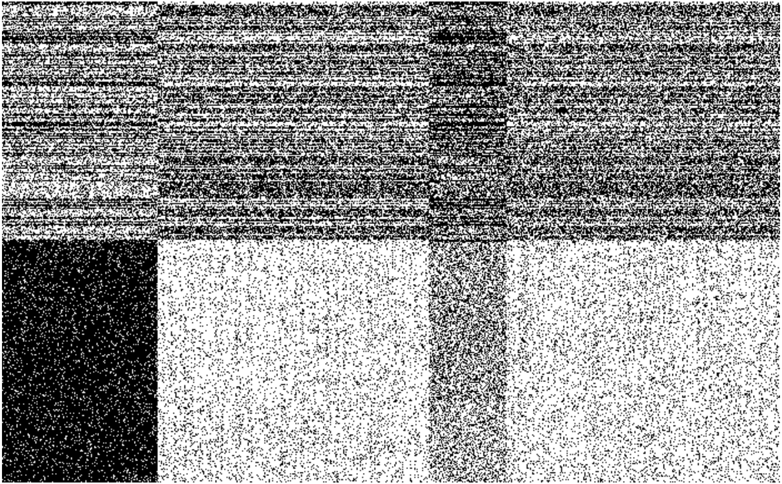
55-fold blow up



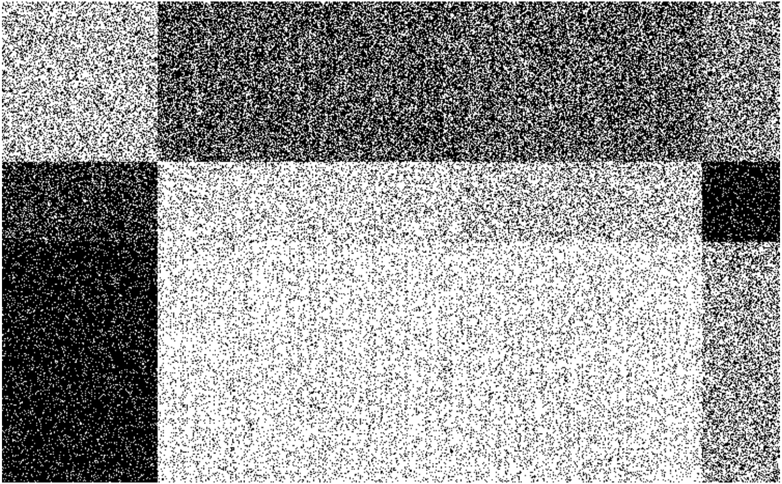
60-fold blow up



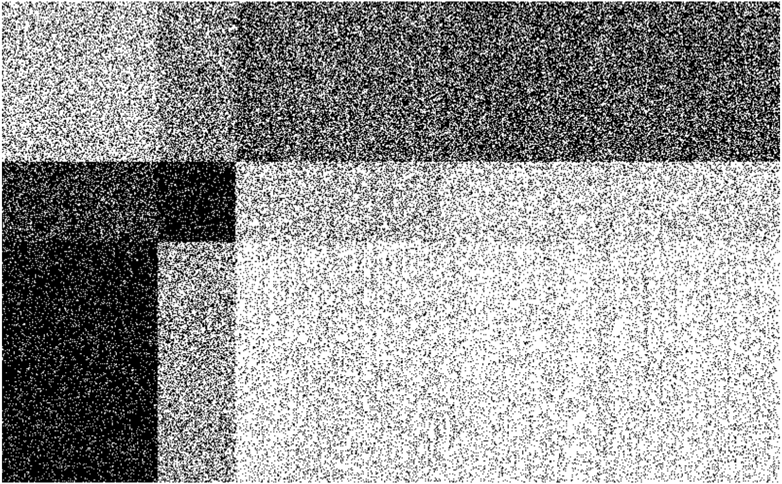
65-fold blow up



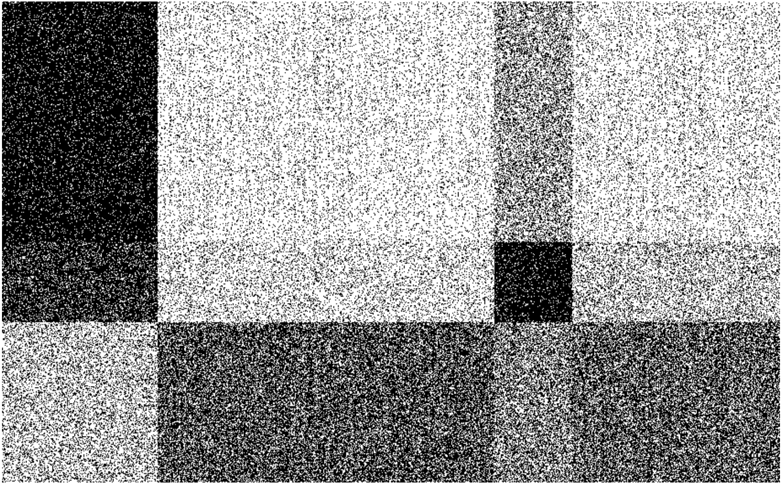
70-fold blow up



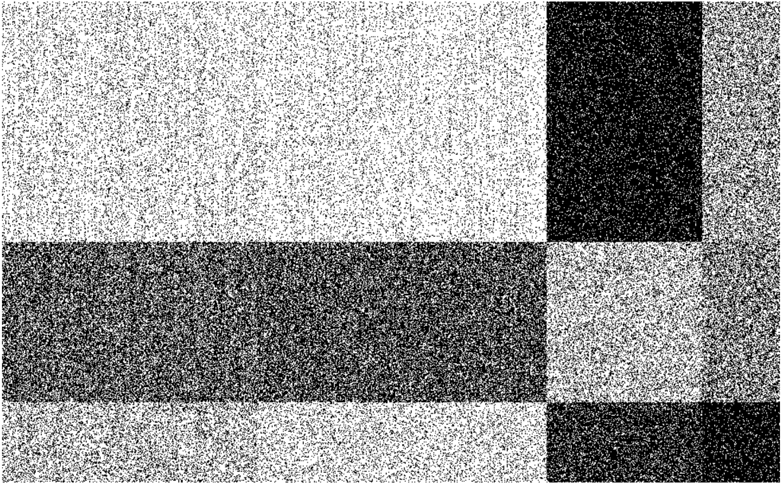
75-fold blow up



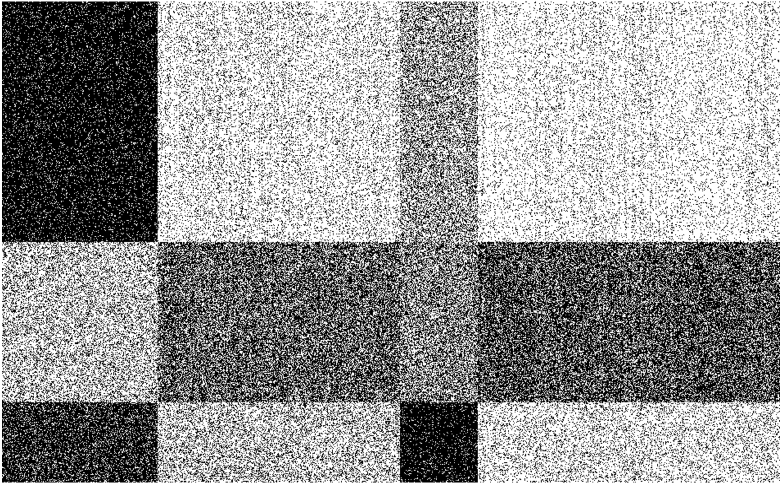
80-fold blow up



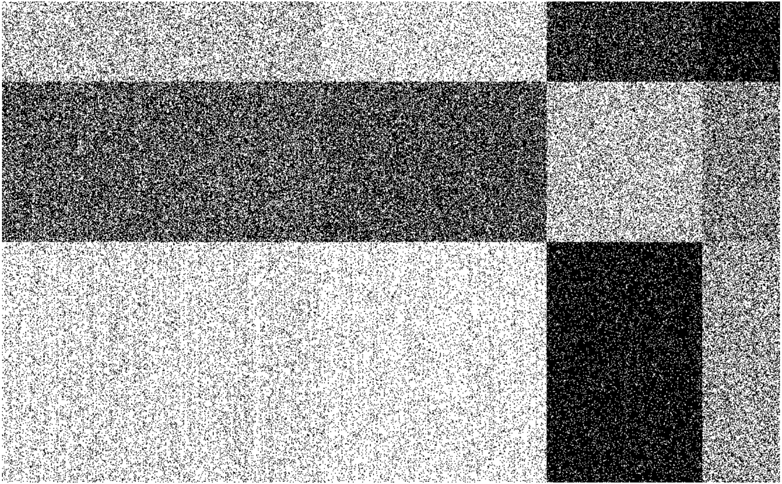
85-fold blow up



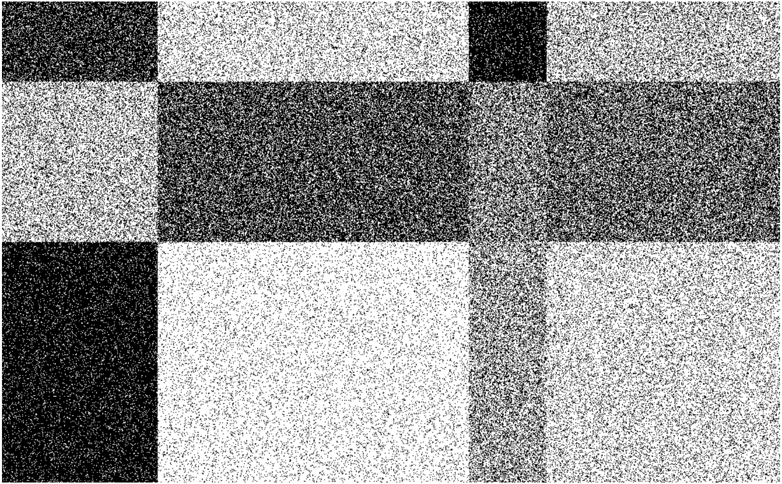
90-fold blow up



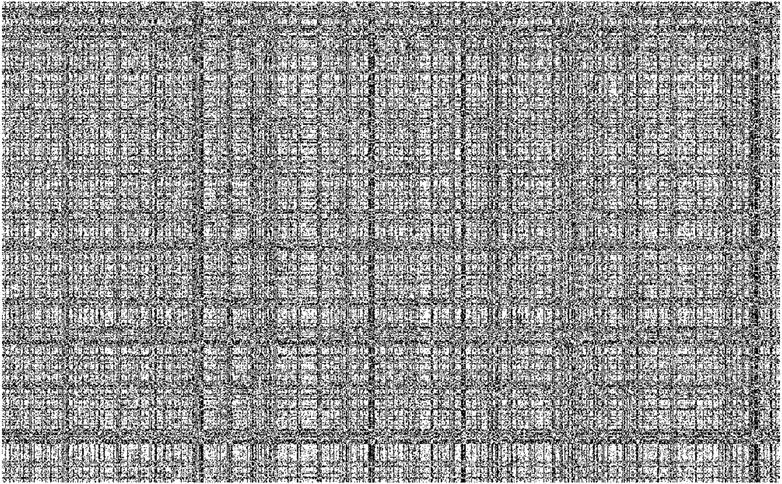
95-fold blow up



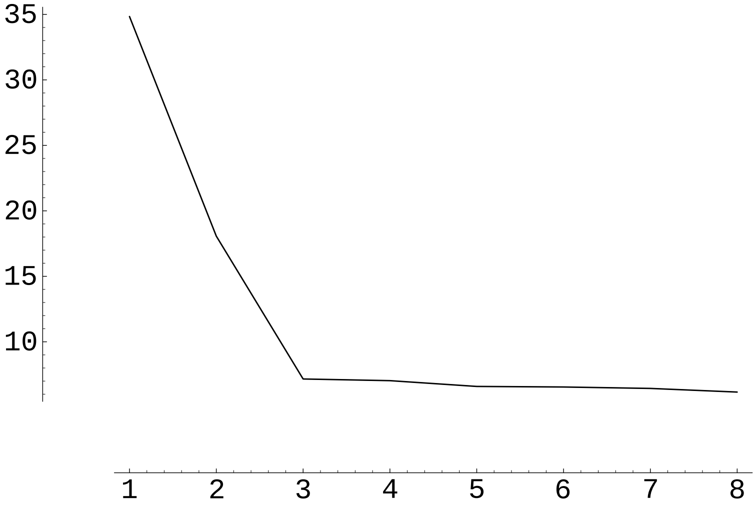
100-fold blow up



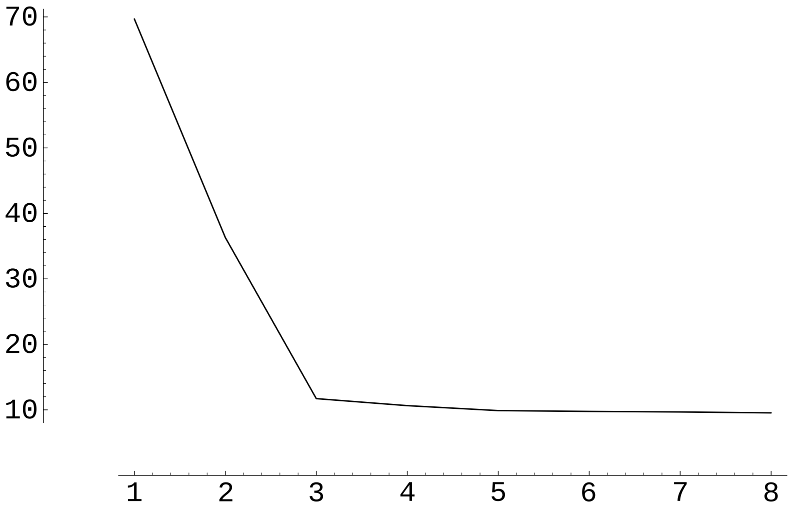
100-fold blow up without sorting



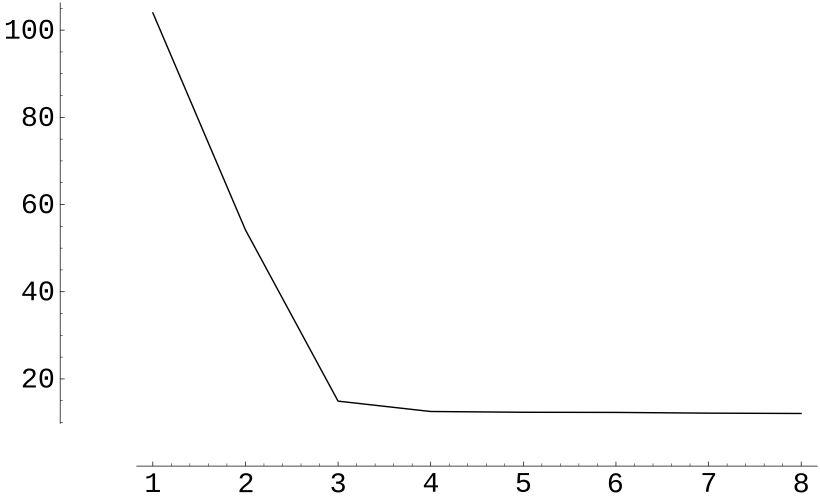
structural singular values (10-fold blow up)



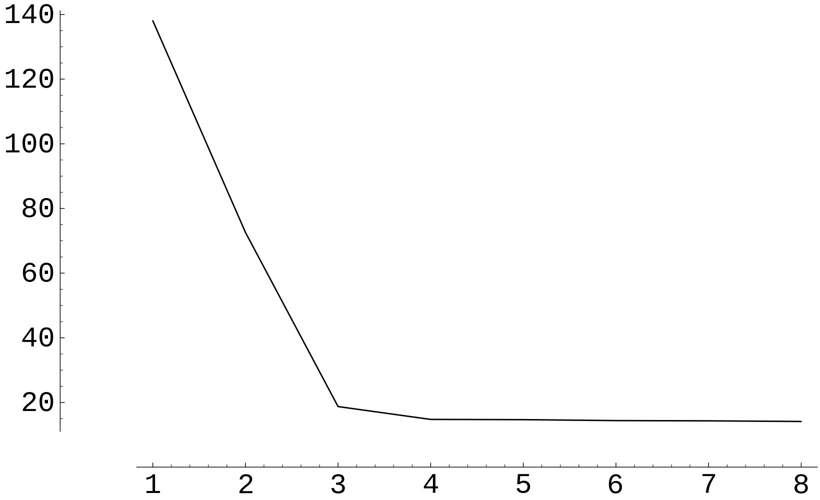
structural singular values (20-fold blow up)



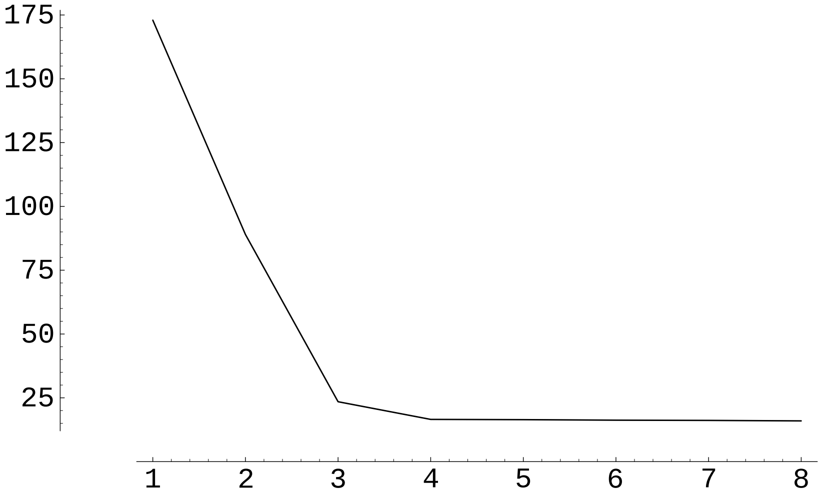
structural singular values (30-fold blow up)



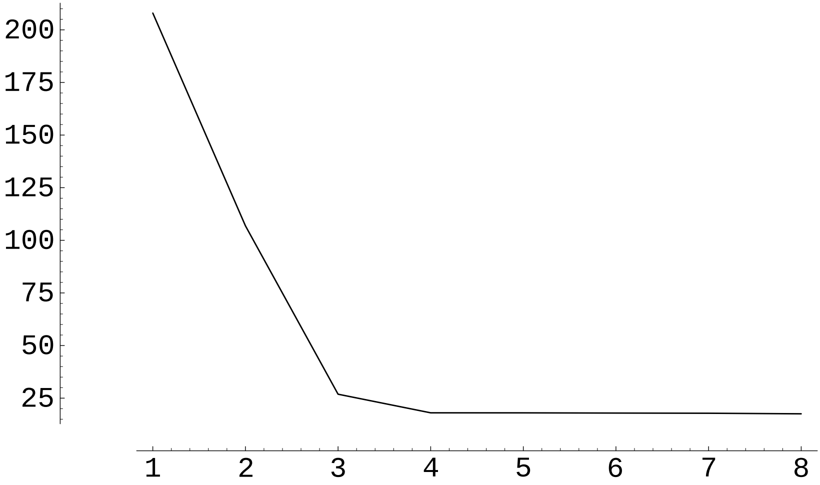
structural singular values (40-fold blow up)



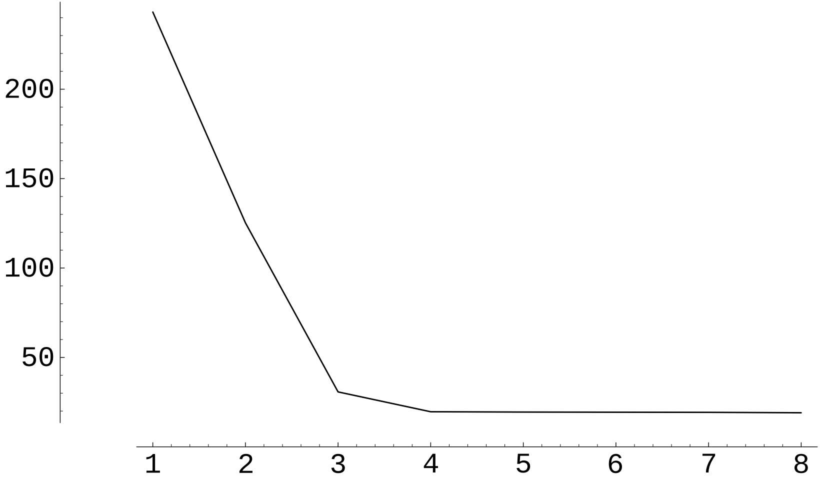
structural singular values (50-fold blow up)



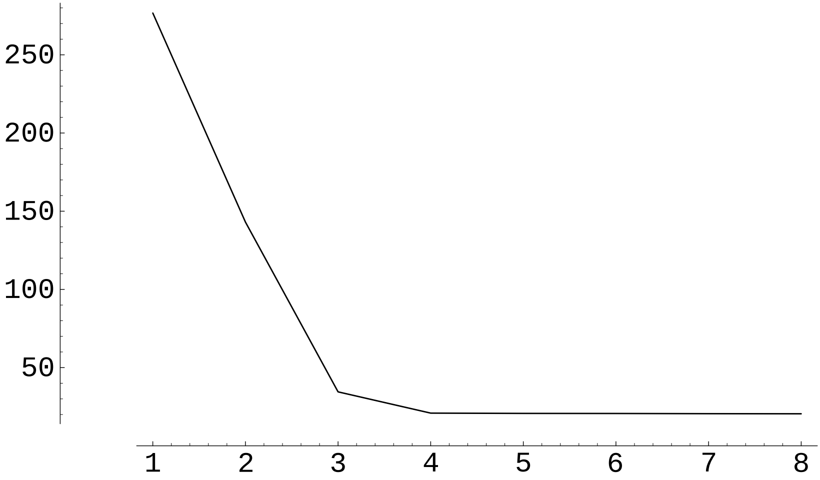
structural singular values (60-fold blow up)



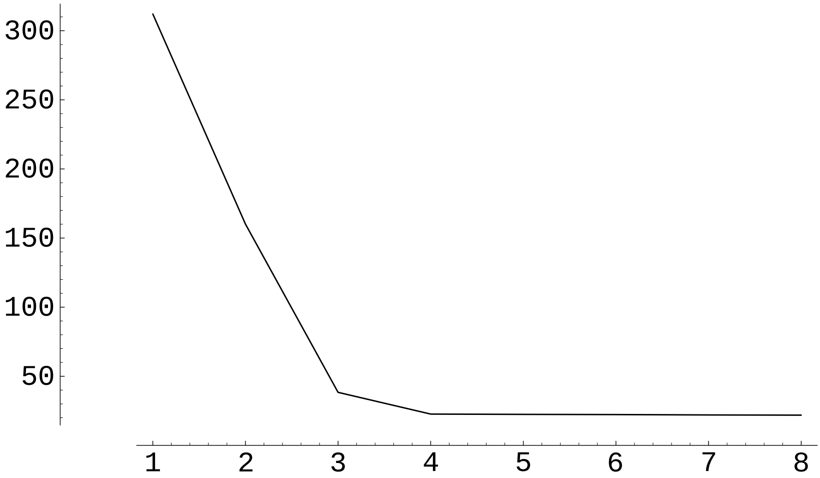
structural singular values (70-fold blow up)



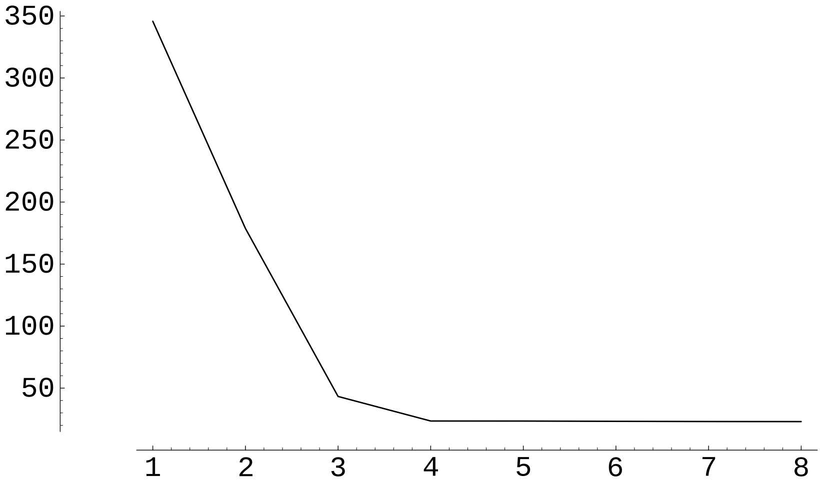
structural singular values (80-fold blow up)



structural singular values (90-fold blow up)



structural singular values (100-fold blow up)



References

- ALDOUS D. J. (1981): Representations for partially exchangeable arrays of random variables. *J. Multivar. Anal. 11, 581-598*.
- BOLLA, M., FRIEDL, K., and KRÁMLI, A. (2010): Singular value decomposition of large random matrices (for two-way classification of microarrays). *J. Multivar. Anal. 101, 434-446*.
- BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T., and VESZTERGOMBI, K. (2008): Convergent sequences of dense graphs I, subgraph frequences, metric properties and testing. *Advances in Mathematics 219, 1801-1851*.
- DIACONIS, P. and Freedman, D. (1981): On the statistics of vision: The Julesz conjecture. *J. Math. Psychol. 24, 112-138*.
- SZEMERÉDI, E. (1978): Regular partitions of graphs. *Proc. of the Colloque Inter. CNRS, 399-401*.