Non-asymptotic theory of random matrices and sparsity

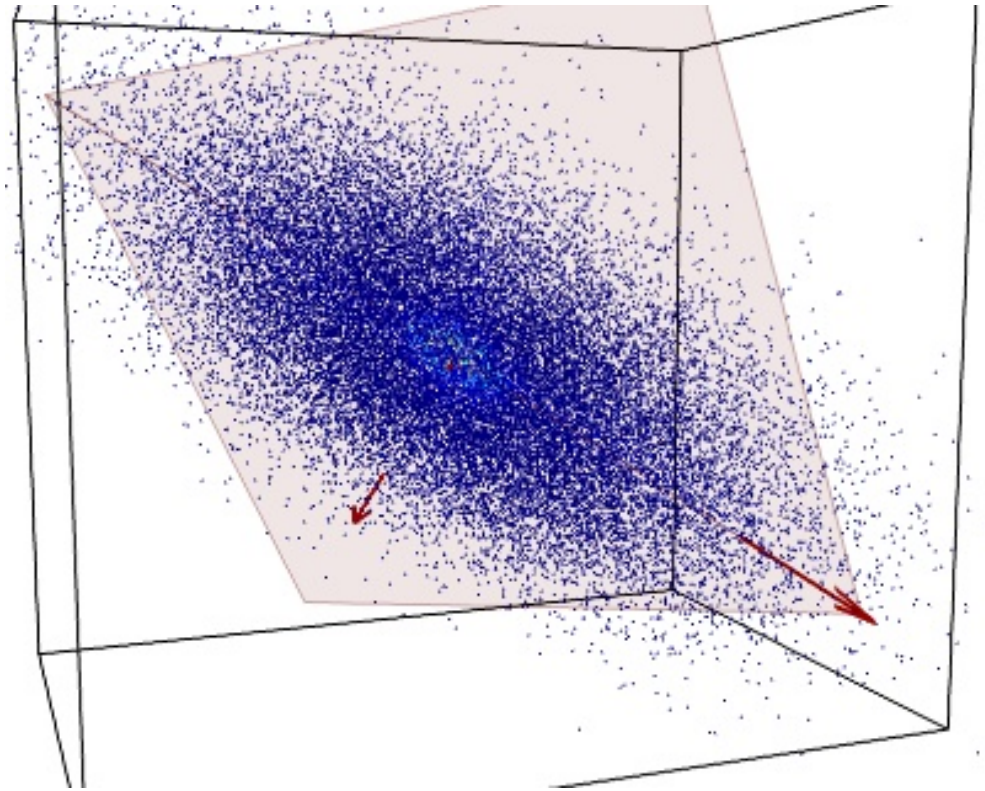
Roman Vershynin

University of Michigan

Sparsity and Computation
Bonn, June 2010

Roots of random matrix theory

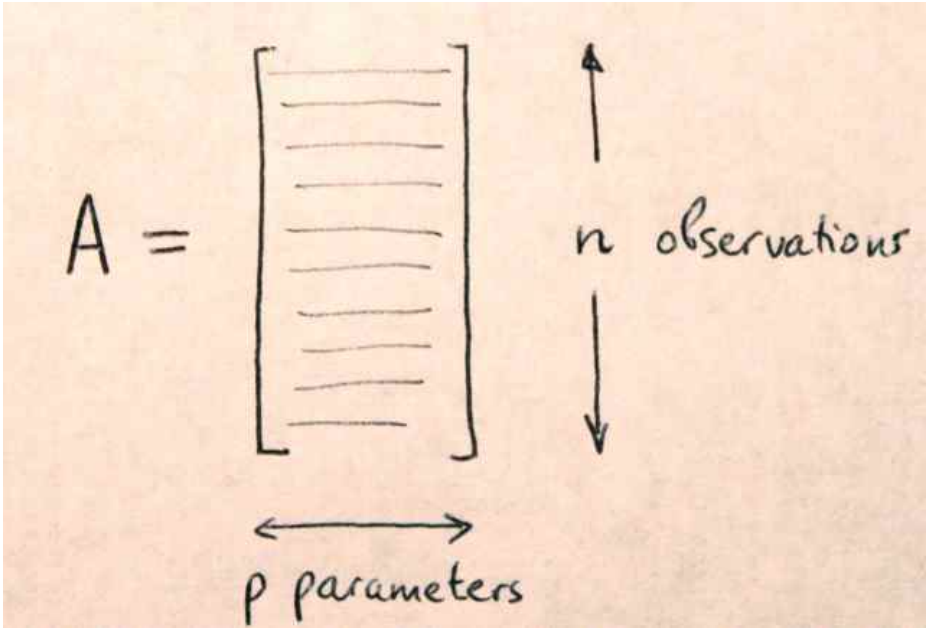
- The need to understand spectra of random matrices.
- Statistics: Principal Component Analysis (PCA) [Wishart'20 ...]



PCA of a multivariate Gaussian distribution. [Gaël Varoquaux's blog gael-varoquaux.info]

Roots of random matrix theory

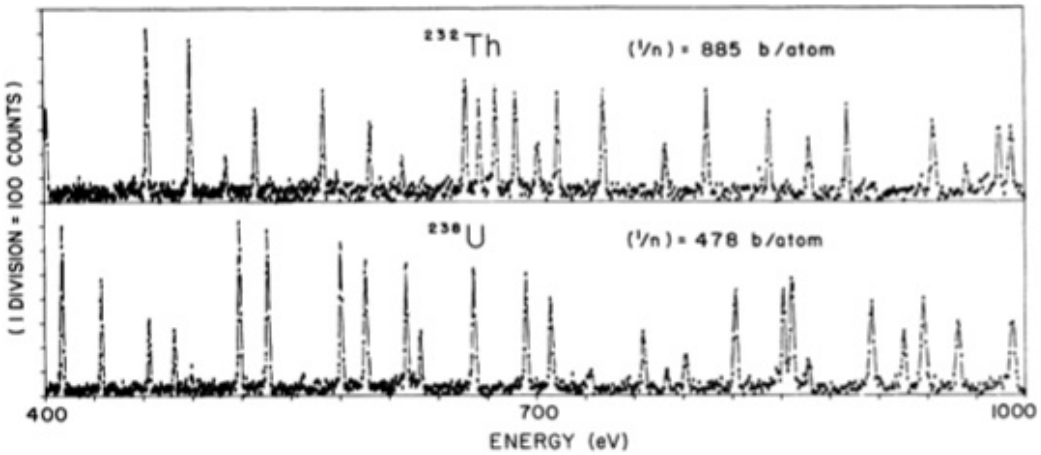
- Collect a sample of n independent points; organize them as rows of the random matrix



- Compute the $p \times p$ Wishart matrix $W = A^T A$. The eigenvalues of $\sqrt{W} = |A|$ are called the singular vectors of A. For the largest singular vectors, the eigenvectors of W are the principal components.
- See [Wikipedia, Estimation of covariance matrices]

Roots of random matrix theory

- Quantum mechanics: excitation spectra of nuclei
  [Wigner, Dyson 50-60...]



Slow neutron resonance on thorium 232 and uranium 238 nuclei [Rahn *et al*, Phys. Rev. C 6 (1972), 1854]

- Distributon different from Poisson: energy levels tend to repel
- Similar to the eigenvalues of Wigner matrices, the $n \times n$ symmetric matrices with i.i.d. entries above the diagonal.
- Heuristics: energy levels are the eigenvalues of a Hamiltonian (complicated, modeled as a random operator)
- See [Mehta, Random matrices (book)]

Roots of random matrix theory

- Numerical analysis: average analysis of matrix algorithms [von Neumann'50, Smale'80...]
- Solvers for systems of linear equations

$$Ax = b$$

- can be tested on random inputs, for $A =$ random matrix.
- Accuracy, speed usually depends on the condition number of A, the ratio of the largest to the smallest singular values:

$$\kappa(A) = \frac{s_{\max}(A)}{s_{\min}(A)}.$$

- Algorithms usually work well for well-conditioned matrices, for which $\kappa(A) = O(1)$ or polynomial in dimension.
- See [Smale'85, Spielman-Teng ICM'02]

Roots of random matrix theory

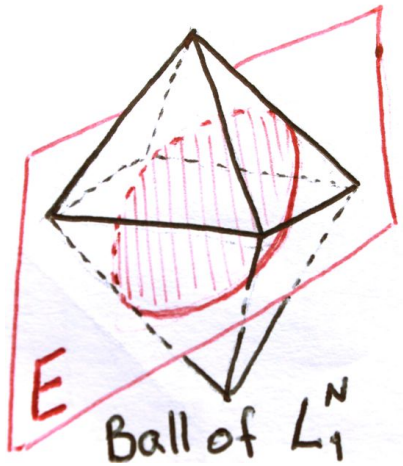
- Functional analysis: probabilistic constructions [Milman, Gluskin'70...]
- In finite dimensional normed spaces $(\mathbb{R}^n, \|\cdot\|)$, random matrices A model "typical" linear operators.
- This is a functional analytic version of the probabilistic method in combinatorics.

Roots of random matrix theory

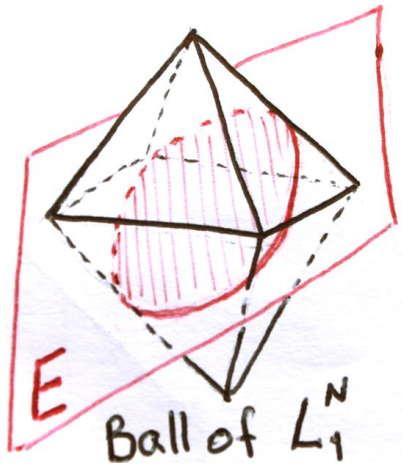
Example: Kashin's theorem on Euclidean subspaces of L_1^N with the norm $\|f\|_{L_1} = \frac{1}{N}\sum_{i=1}^N |f(i)|$.

Theorem (Euclidean subspaces) [Kashin'77].
For all $N = (1+\delta)n$, there exist subspace E of L_1^N of dimension n which is uniformly isomorphic to L_2^n:

$$\|f\|_{L_2} \lesssim_\delta \|f\|_{L_1} \le \|f\|_{L_2} \quad \text{for all } f \in E.$$



Roots of random matrix theory



- Kashin constructs E as a kernel of a random Bernoulli matrix with independent ± 1 entries. Other constructions possible: image instead of kernel [Litvak et al'05, Rudelson'06]; random Gaussian matrix with independent $N(0,1)$ entries, random orthogonal matrix uniform in $O(n)$. Spaces more general than L_1 also possible (with the same volumetric properties) [Szarek-Tomczak'80, Litvak et al'05].
- Open problem: deterministic constructions.
- See [Any of V. Milman's surveys]

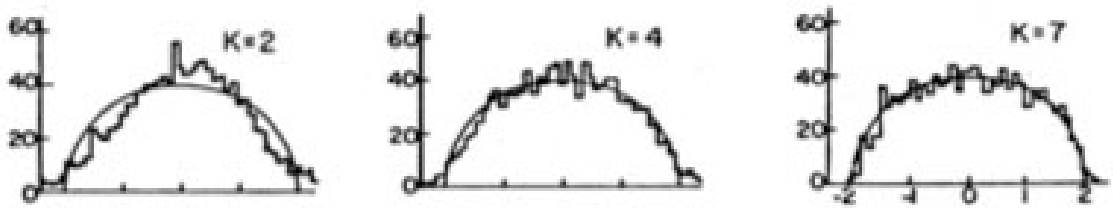
Limit laws in random matrix theory

Wigner's semicircle law '58. Consider an $n \times n$ symmetric Gaussian matrix A_n, whose above diagonal entries are independent $N(0,1)$. As $n \to \infty$, the spectrum of $\frac{1}{\sqrt{n}} A_n$ is distributed according to the semicircle law with density

$$\frac{1}{2\pi} \sqrt{4 - x^2} \quad \text{on } [-2, 2].$$

Precisely, if $S_n(z)$ is the empirical spectral distribution function of $\frac{1}{\sqrt{n}} A_n$ (the number of eigenvalues $\leq z$), then

$$\frac{S_n(z)}{n} \to \frac{1}{2\pi} \int_{-\infty}^{z} (4 - x^2)_+^{1/2} \, dx \quad \text{almost surely as } n \to \infty.$$



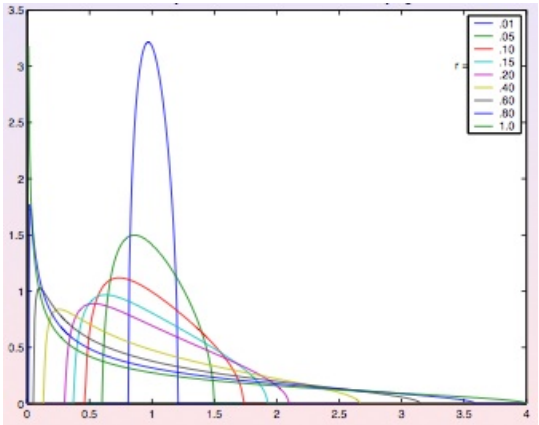
[J. French, S. Wong, Phys. Lett. B. 35 (1971), 5]

Limit laws in random matrix theory

- Marchenko-Pastur law governs the limiting spectrum of $n \times n$ Wishart matrices $W_{N,n} = A^T A$, where $A = A_{N,n}$ is an $N \times n$ *random Gaussian matrix* with i.i.d. $N(0,1)$ entries.

Marchenko-Pastur law '67. As the dimensions $N, n \to \infty$ while aspect ratio $n/N \to y \in (0,1]$, the spectrum of $\frac{1}{N} W_{N,n}$ has limiting density

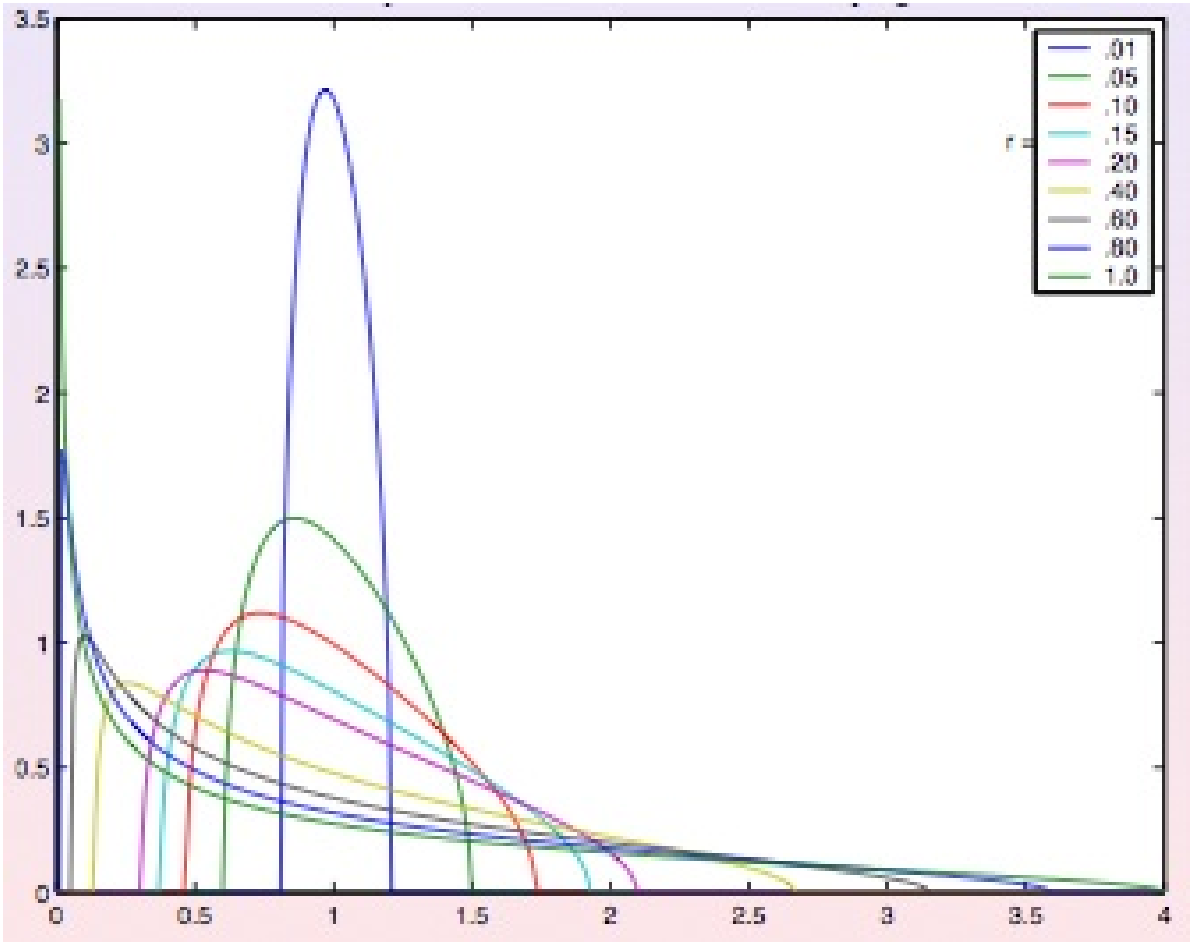
$$\frac{1}{2\pi xy} \sqrt{(b-x)(x-a)} \quad \text{where } a = (1-\sqrt{y})^2, b = (1+\sqrt{y})^2.$$



[El Karoui, Estimation of large dimensional sparse covariance matrices, 2009]

Limit laws in random matrix theory

Limiting density in Marchenko-Pastur law:

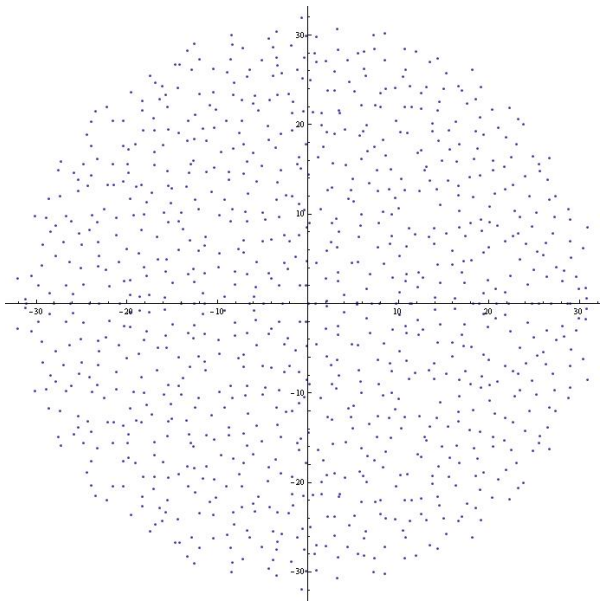


[El Karoui, Estimation of large dimensional sparse covariance matrices, 2009]

Limit laws in random matrix theory

- Circular law governs the limiting spectrum of $n \times n$ random Gaussian matrices A_n with i.i.d. $N(0,1)$ entries.

Circular law [Mehta'67]. As the dimension $n \to \infty$ the spectrum of $\frac{1}{\sqrt{n}} A_n$ is distributed according to the uniform measure on the unit disc $\{z \in \mathbb{C} : |z| = 1\}$.



[B.Valkó, A course on random matrices, math.wisc.edu/~valko/courses/833/833.html]

Universality

- It is widely believed that phenomena typically observed in statistical physics and in asymptotic random matrix theory are universal – independent of the distribution of the entries.
- Analogy with classical probability: for independent $N(0,1)$ random variables Z_i, their normalized sum

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i$$

  is again a standard normal random variable.
- Central Limit Theorem: the above fact is universal as $n \to \infty$. Valid for general i.i.d. Z_i with zero mean and unit variance.
- Universality in random matrix theory: holds!
  - Wigner's semicircle law [Pastur'73, Bai-Silverstein'10]
  - Marchenko-Pastur law [Watcher'78, see Bai'99]
  - circular law [Girko'84, Bai'97, Götze-Tikhomirov'08, Tao-Vu'08-10]

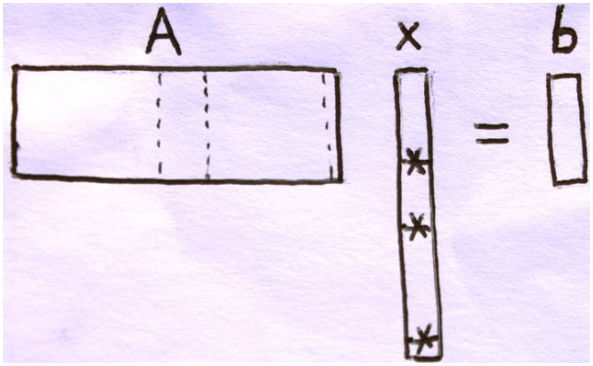
Asymptotic and non-asymptotic regimes

- Asymptotic random matrix theory offers remarkable predictions as dimensions grow to infinity. This fits very well the purposes of statistical physics.
- However, there is often lack of understanding of finite (fixed but large) dimensions. Many applications operate there:
  - statistics (number of parameters is fixed),
  - numerical analysis of algorithms (number of variables and equations is fixed),
  - functional analysis (operators act on fixed spaces).
- Asymptotic regime = dimensions grow to infinity; precise limiting phenomena
- Non-asymptotic regime = any fixed dimensions; results are optimal up to constants

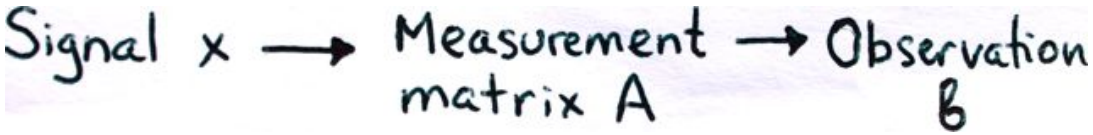
Compressed Sensing

New connections arise in compressed sensing.

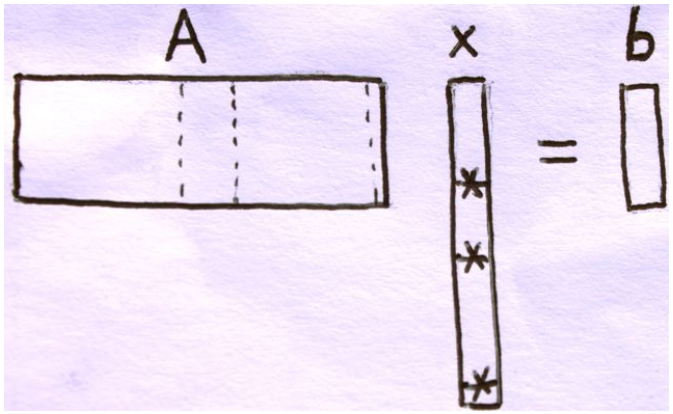
Sparse recovery problem. Find the sparsest solution to an underdetermined system of linear equations $Ax = b$ where A is an $m \times N$ matrix, $m \ll N$.



Many applications. x: unknown sparse signal, A: known observation matrix, $b = Ax$: known observation of x. Problem: recover the signal x from the observation b.



Compressed Sensing



- In other words, we want to solve the optimization problem

$$\min \|x\|_0 \text{ subject to } Ax = b$$

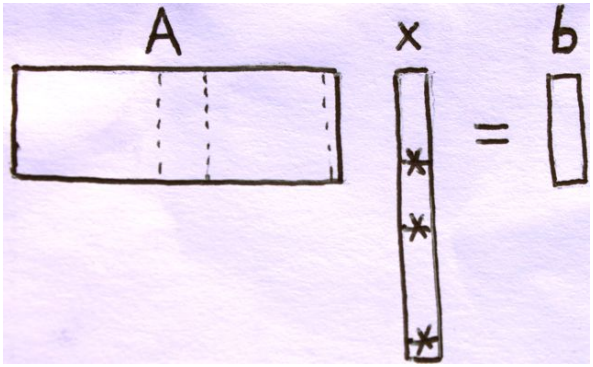
where $\|x\|_0 = |\operatorname{supp}(x)| =$ number of nonzero coordinates.

- We relax this non-convex problem to the convex program

$$\min \|x\|_1 \text{ subject to } Ax = b.$$

- When are these two problems equivalent?

Compressed Sensing



Answer: when the operator A respects all sparse vectors.

Definition. A matrix A is a restricted isometry given sparsity s if

$$(1-\delta)\|x\|_2 \le K\|Ax\|_2 \le (1+\delta)\|x\|_2 \quad \text{for all } x, \|x\|_0 \le s$$

where K is any normalization factor, and $\delta = \delta_s \in (0,1)$ is small.

Theorem [Candes, Tao'04]. Assume that an s-sparse vector x is a solution to $Ax = b$, where A is a restricted isometry with $\delta_{2s} \le 0.2$. Then x can be recovered from b by the convex program

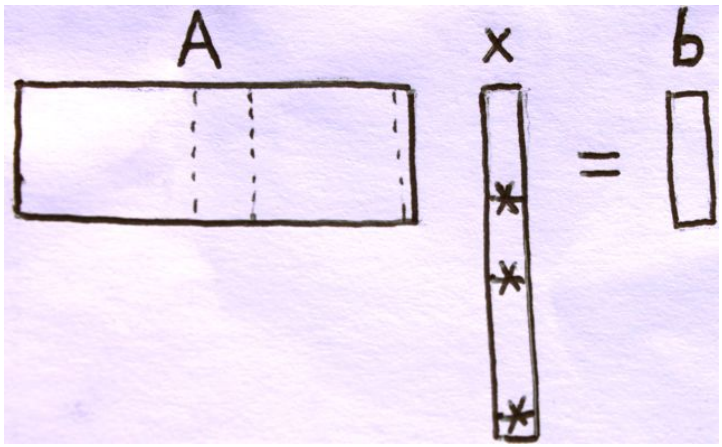
$$\min \|x\|_1 \text{ subject to } Ax = b.$$

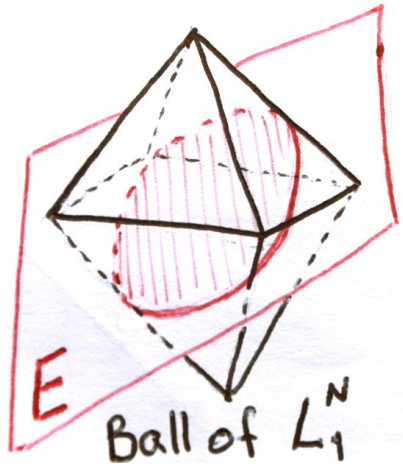
Restriced isometries

Restricted isometries form an interesting class of matrices. They have been implicitly used in geometric functional analysis in constructions of Euclidean subspaces:

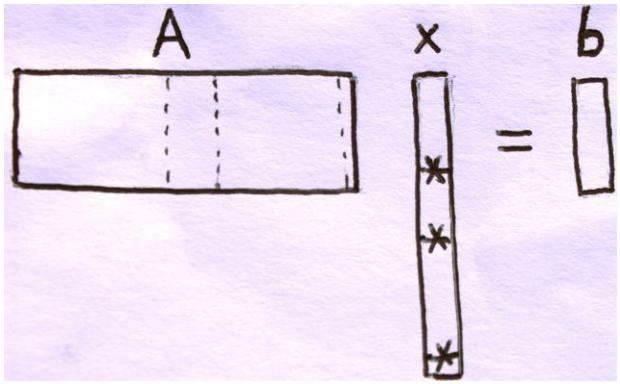
Observation (Restricted isometries imply Kashin-type sections). Assume A is an $m \times N$ matrix with orthonormal rows, $m = \varepsilon N$. If A is a restricted isometry with $\delta = \delta_{\gamma N} < 1$, then A^T acts as an almost isometric embedding of L_2^m into L_1^N:

$$\|f\|_{L_2} \lesssim_{\varepsilon,\gamma} \|A^T f\|_{L_1} \le \|f\|_{L_2} \text{ for all } f \in L_2^m.$$



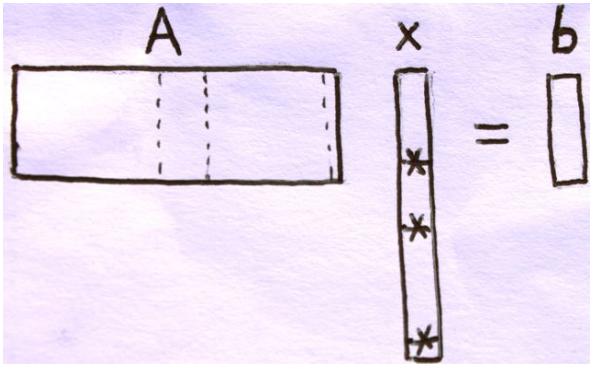


Restriced isometries



- Connection with random matrix theory: Restricted isometries are difficult to construct. Deterministic constructions with good dimensions are uknown (e.g. those yielding Kashin's sections of proportional dimensions).
- The best known constructions of restricted isometries are randomized, i.e. random matrices:

Restriced isometries

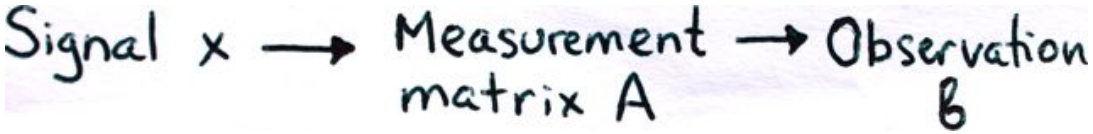


Theorem (Random matrices are restricted isometries). Let A be an $m \times N$ Gaussian random matrix (entries $= \mathcal{N}(0,1)$ iid). If

$$m \sim s \log(N/s)$$

then with high probability A is a restricted isometry for s-sparse vectors, with $\delta_s \leq 0.1$.

This result allows one to recover a sparse signal from few measurements ($m \sim$ sparsity s, not the dimension N).



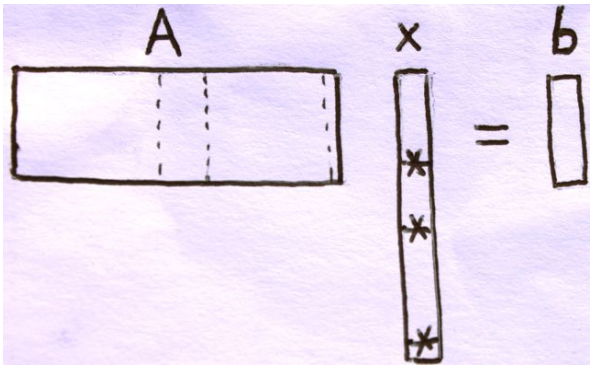
Restriced isometries

General measurement ensembles:

- The same result holds for Bernoulli random matrices (± 1 entries), and generally for subgaussian random matrices.
- A similar result holds for partial Fourier random matrices, which are obtained by randomly selecting m rows from an $N \times N$ Discrete Fourier Transform matrix. However, the number of measurements is a bit higher: $m \sim s \log^4 N$ [Candes-Tao, Rudelson-Vershynin].

Approximate isometries

Theorem (Random matrices are restricted isometries). Let A be an $m \times N$ Gaussian random matrix. If $m \sim s \log(N/s)$ then with high probability A is a restricted isometry for s-sparse vectors.



- Why true? Because a given $n \times s$ random Gaussian matrix A_I is an approximate isometry with very high probability (allowing one to take the union bound over all submatrices A_I of A):
- Indeed, A_I is tall, $n \ll s$, thus we have with high probability

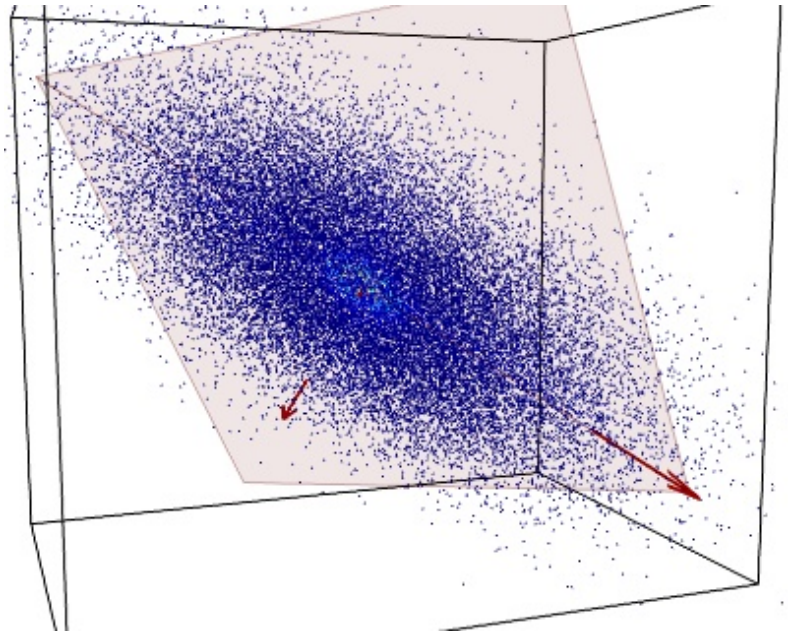
$$\|A_I x\|_2 \approx \|x\|_2 \text{ for all } x \in \mathbb{R}^s.$$

This is proved using concentration of measure in Gauss space.

Approximate isometries in statistics

Question. When a tall random matrix is an approximate isometry?

- This question brings us back to a basic problem in statistics – estimating the covariance structure of a high dimensional distribution.



PCA of a multivariate Gaussian distribution. [Gaël Varoquaux's blog `gael-varoquaux.info`]

Covariance matrix

- The covariance structure of a centered high-dimensional distribution μ (equivalently, a random vector $\mathbf{X}$ distributed according to μ) is captured by its covariance matrix

$$\Sigma = \mathbb{E}\,\mathbf{X}\mathbf{X}^T = (\mathbb{E}X_iX_j)_{i,j=1}^p = (\mathrm{cov}(X_i, X_j))_{i,j=1}^p$$

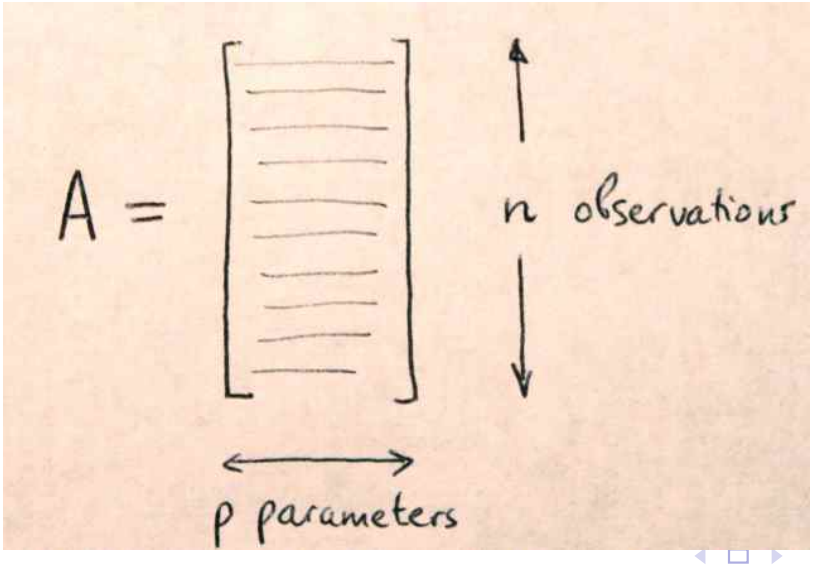
- $\Sigma = \Sigma(\mathbf{X})$ is a symmetric, positive semi-definite $p \times p$ matrix. It is a multivariate version of the variance $\mathrm{Var}(X)$.
- If $\Sigma(\mathbf{X}) = I$ we say that $\mathbf{X}$ is isotropic. Every full dimensional random vector $\mathbf{X}$ can be made into an isotropic one by the linear transformation $\Sigma^{-1/2}\mathbf{X}$.

Estimation of covariance matrices

- Problem: estimate Σ. It arises in signal processing, genomics, financial mathematics, pattern recognition, convex geometry.
- We take a sample of n independent points $\mathbf{X}_1, \dots, \mathbf{X}_n$ from the distribution. We hope to estimate Σ by the sample covariance matrix

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_k \mathbf{X}_k^T = \frac{1}{n} A^T A,$$

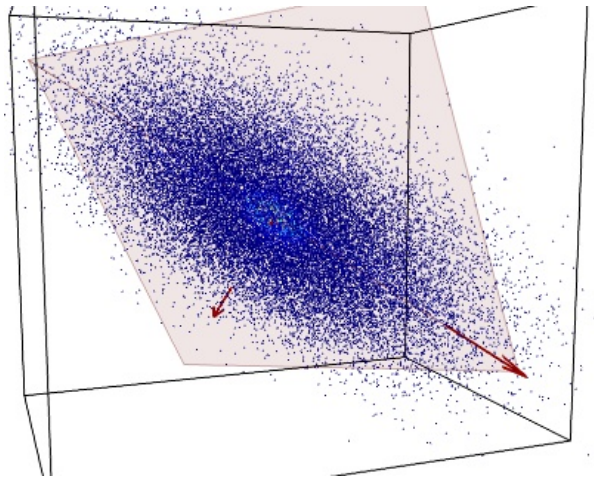
where A is a tall random matrix with independent rows:



Estimation of covariance matrices

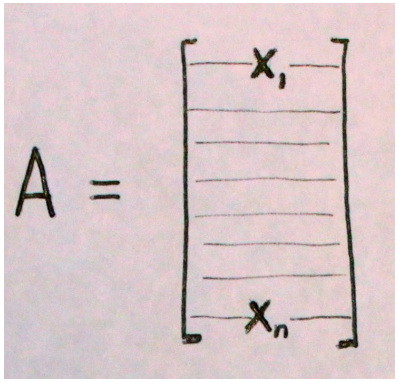
Covariance Estimation Problem. Determine the minimal sample size $n = n(p)$ that guarantees with high probability (say, 0.99) that the sample covariance matrix Σ_n estimates the actual covariance matrix Σ with fixed accuracy (say, $\varepsilon = 0.01$) in the operator norm:

$$\|\Sigma_n - \Sigma\| \le \varepsilon \|\Sigma\|.$$



PCA of a multivariate Gaussian distribution. [Gaël Varoquaux's blog `gael-varoquaux.info`]

Estimation problem and random matrices



$$\Sigma_n = \frac{1}{n} A^T A.$$

- For isotropic distributions ($\Sigma = I$), the desired estimation $\|\Sigma_n - I\| \le \varepsilon$ is equivalent to saying that $\frac{1}{\sqrt{n}} A$ is an approximate isometry:

$$(1-\varepsilon)\sqrt{n} \le \|Ax\|_2 \le (1+\varepsilon)\sqrt{n} \quad \text{for all } x \in S^{p-1}.$$

- Equivalently, the singular values $s_i(A) = \mathrm{eig}(A^T A)^{1/2}$ are all close to each other and to $\sqrt{n}$:

$$(1-\varepsilon)\sqrt{n} \le s_{\min}(A) \le s_{\max}(A) \le (1+\varepsilon)\sqrt{n}.$$

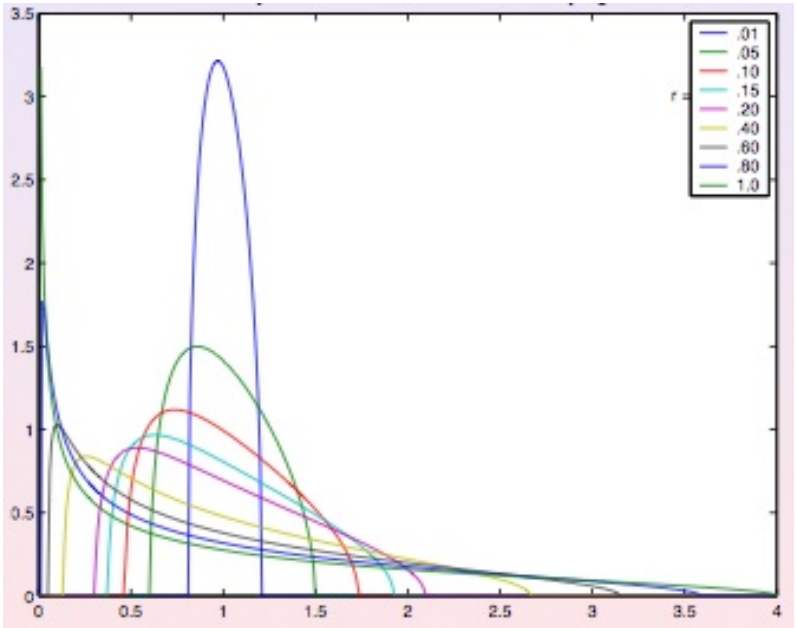
Question. What random matrices with independent rows are approximate isometries?

Random matrices with independent entries

- Simplest example: Gaussian distributions.
  A is a $p \times n$ random matrix with independent $N(0,1)$ entries.
- The endpoints of Marchenko-Pastur law suggest that, as $n, p \to \infty$, $n/p \to$ const, we have

$$s_{\min}(A) \to \sqrt{n} - \sqrt{p}, \quad s_{\max}(A) \to \sqrt{n} + \sqrt{p} \quad \text{a.s.}$$

- This is indeed true: Bai-Yin law [+Silverstein, Krishnaiah].



Random matrices with independent entries

Bai-Yin: $s_{\min}(A) \to \sqrt{n} - \sqrt{p}$, $s_{\max}(A) \to \sqrt{n} + \sqrt{p}$.

- Thus making n slightly bigger than p we force both extreme values to be close to each other, and make A an almost isometry.
- This easily translates into the statement that the sample covariance matrix $\Sigma_n = \frac{1}{n} A^T A$ nicely approximates the actual covariance matrix I:

$$\|\Sigma_n - I\| \approx 2\sqrt{\frac{p}{n}} + \frac{p}{n}.$$

Answer to the Estimation Problem for Gaussian distributions.
Sample size $n \sim p$ suffices to estimate the covariance matrix by the sample covariance matrix: $\|\Sigma_n - \Sigma\| \le \varepsilon \|\Sigma\|$.

Non-gaussian distribtions

Answer to the Estimation Problem for Gaussian distributions.
Sample size $n \sim p$ suffices to estimate the covariance matrix in $\mathbb{R}^p$ by the sample covariance matrix: $\|\Sigma_n - \Sigma\| \le \varepsilon \|\Sigma\|$.

- The same answer holds for more general distributions:
- OK for sub-gaussian distributions, whose one-dimensional marginals have tails dominated by gaussian, $\mathbb{P}\{|\langle \mathbf{X}, x\rangle| \ge t\} \le 2\exp(-ct^2)$. Reason: ϵ-net argument.
- OK for sub-exponential distributions, those with heavier tails $2\exp(-ct)$ [Adamczak, Litvak, Pajor, Tomczak'09].
- OK for distributions with finite 4-th moments, up to a $\log\log p$ factor. [V'10]. Conjecture: $\log\log p$ is not needed.
- OK for arbitrary distributions up to a $\log p$ moment [Rudelson'99], [Bourgain'99]. Reason: operator-valued deviation inequalities for sums of independent random matrices, see e.g. [Tropp'10].

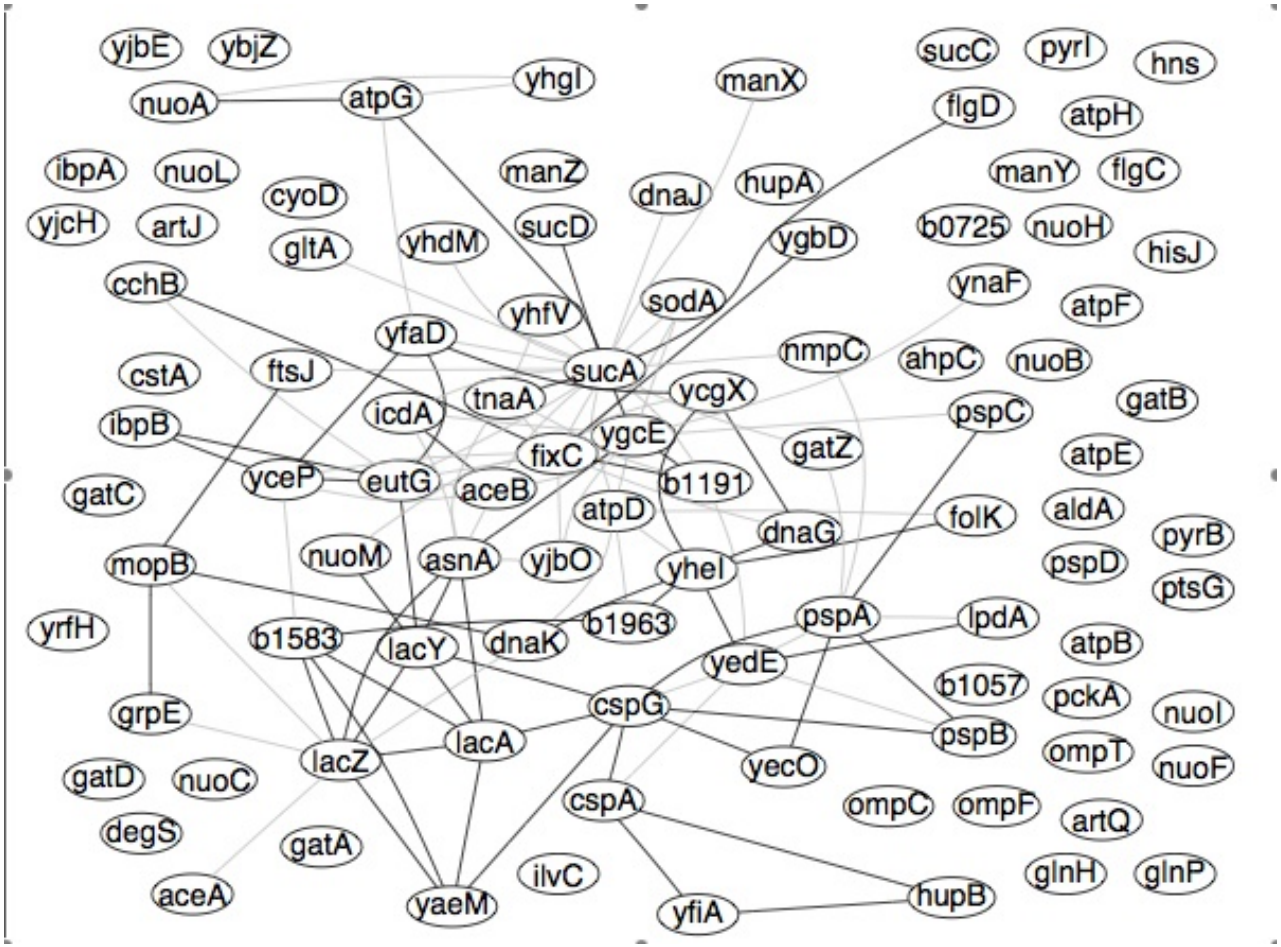
Sparse estimation of covariance matrices

- Like in compressed sensing, most of today's practical applications require very small sample sizes n compared with the dimension p (number of parameters), calling for

$$n \ll p.$$

- In this regime, covariance estimation is generally impossible for dimension reasons. But usually (in practice) one knows a priori some structure of the covariance matrix Σ.
- For example, as in compressed sensing settng, Σ may be known to be sparse, having few non-zero entries (i.e. most random variables are uncorrelated). Example:

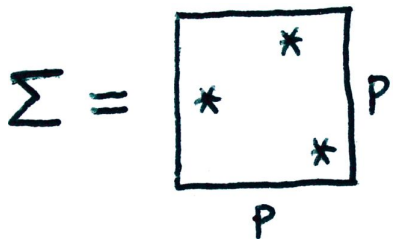
Covariance graph



Gene association network of E. coli [J. Schäfer, K. Strimmer'05]

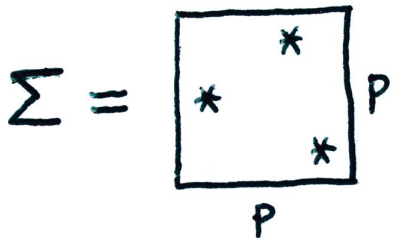
Sparse estimation of covariance matrices

Sparse Estimation Problem. Consider a distribution in $\mathbb{R}^p$ whose covariance matrix Σ has at most $s \le p$ nonzero entries in each column (equivalently, each component of the distribution is correlated with at most s other components). Determine the minimal sample size $n = n(p, s)$ needed to estimate Σ with a fixed error in the operator norm, and with high probability.



A variety of techniques has been proposed in statistics, notably the shrinkage methods going back to Stein.

Sparse estimation of covariance matrices



- The sparse estimation problem is nontrivial even for Gaussian distributions, and even if we know the locations of the non-zero entries of Σ. Let's assume this (otherwise take the biggest entries of Σ_n).

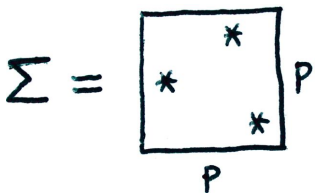
Method for sparse covariance estimation: Take the sample of n points and compute the sample covariance matrix Σ_n. Zero out the entries that are known to be zero a priori. The resulting sparse matrix should be a good estimator for Σ.

- Zeroing out amounts to taking Hadamard product (entrywise) $M \cdot \Sigma_n$ with a given sparse 0/1 matrix M (mask).
- Does this method work? Yes:

Sparse estimation of covariance matrices

Theorem (Sparse Estimation) [Levina-V'10]
Consider a centered Gaussian distribution in $\mathbb{R}^p$ with covariance matrix Σ. Let M be a symmetric $p \times p$ "mask" matrix with $0, 1$ entries and with at most s nonzero entries in each column. Then

$$\mathbb{E}\|M \cdot \Sigma_n - M \cdot \Sigma\| \le C \log^3 p \left(\sqrt{\frac{s}{n}} + \frac{s}{n} \right) \cdot \|\Sigma\|.$$



Compare this with the consequence of the Bai-Yin law:

$$\mathbb{E}\|\Sigma_n - \Sigma\| \approx \left(2\sqrt{\frac{p}{n}} + \frac{p}{n} \right) \|\Sigma\|.$$

This matches the Theorem in the non-sparse case $s = p$.

Corollary. Sample size $n \sim s \log^6 p$ suffices for sparse estimation.

Sparse estimation of covariance matrices

More generally:

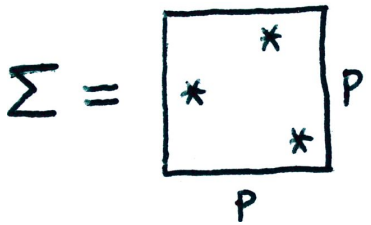
Theorem (Estimation of Hadamard Products) [Levina-V'10]
Consider a centered Gaussian distribution on $\mathbb{R}^p$ with covariance matrix Σ. Then for every symmetric $p \times p$ matrix M we have

$$\mathbb{E}\|M \cdot \Sigma_n - M \cdot \Sigma\| \le C \log^3 p \Big(\frac{\|M\|_{1,2}}{\sqrt{n}} + \frac{\|M\|}{n} \Big) \cdot \|\Sigma\|.$$

where $\|M\|_{1,2} = \max_j (\sum_i m_{ij}^2)^{1/2}$ is the $\ell_1 \to \ell_2$ operator norm.

- This result is quite general. Applies for arbitrary Gaussian distributions (no covariance structure assumed), arbitrary mask matrices M.

Sparse estimation: location of support



Question. So far we assumed that we knew the locations of nonzero elements of the covariance matrix Σ. If we did not (similarly to compressed sensing), how to find them?

- Thresholding: we can just choose the few biggest elements of the sample covariance matrix Σ_n.
- But this requires the entry-wise approximation of Σ by Σ_n. It is OK if $n \gtrsim h^{-2}$ where h is a lower bound on the entries of Σ.
- Is thresholding the best way to estimate the location of nonzeros?

References

- Survey: M. Rudelson, R. Vershynin, *Non-asymptotic theory of random matrices: extreme singular values*, 2010.
- Marginal Estimation: R. Vershynin, *Approximating the moments of marginals of high dimensional distributions*, 2009.
- Covariance Estimation: R. Vershynin, *How close is the sample covariance matrix to the actual covariance matrix?* 2010.
- Sparse Covariance Estimation: L. Levina, R. Vershynin, *Sparse estimation of covariance matrices*, in progress, 2010.