Estimation of covariance matrices

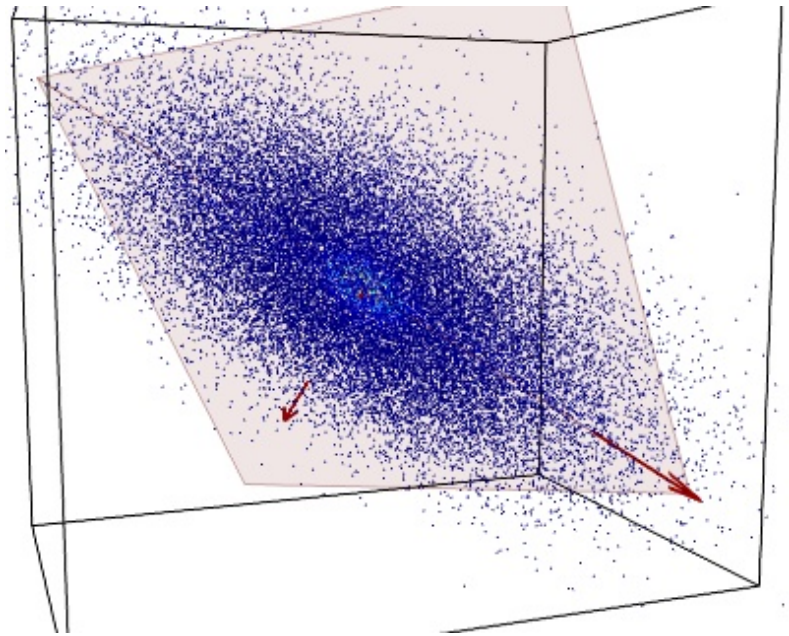
Roman Vershynin

University of Michigan

Probability and Geometry in High Dimensions
Paris, May 2010

Covariance matrix

- Basic problem in multivariate statistics:
  by sampling from a high-dimensional distribution, determine its covariance structure.
- Principal Component Analysis (PCA): detect the principal axes along which most dependence occurs:



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

Covariance matrix

- The covariance structure of a high-dimensional distribution μ is captured by its covariance matrix Σ.
- Let $\mathbf{X}$ be a random vector in $\mathbb{R}^p$ distributed according to μ. We may assume that $\mathbf{X}$ is centered (by estimating and subtracting $\mathbb{E}\mathbf{X}$). The covariance matrix of $\mathbf{X}$ is defined as

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

- $\Sigma = \Sigma(\mathbf{X})$ 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

- Estimation of covariance matrices is a basic problem in multivariate statistics. It arises in signal processing, genomics, nancial mathematics, pattern recognition, computational convex geometry.
- We take a sample of n independent points $\mathbf{X}_1, \dots, \mathbf{X}_n$ from the distribution and form the sample covariance matrix

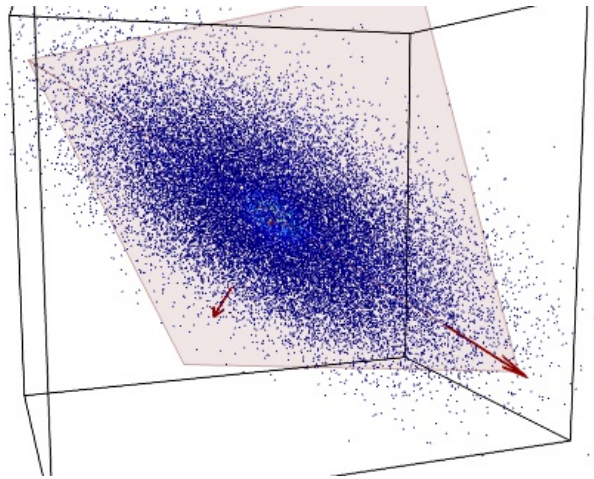
$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_k \mathbf{X}_k^T.$$

Σ_n is a random matrix. Hopefully it approximates Σ well:

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:

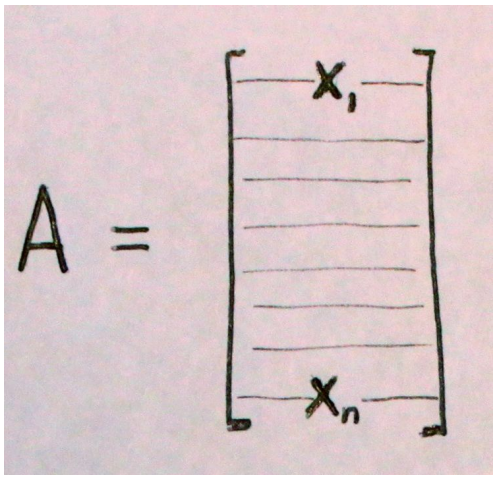
$$\|\Sigma_n - \Sigma\| \le \varepsilon \|\Sigma\|.$$



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

Estimation problem and random matrices

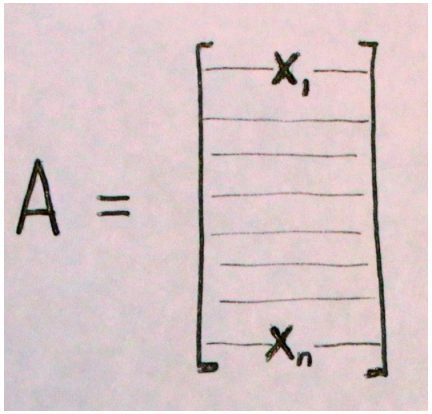
- Estimation problem can be stated as a problem on the spectrum of random matrices.
- Assume for simplicity that the distribution is isotropic, $\Sigma = I$.
- Form our sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ into a $n \times p$ random matrix A with independent rows:



- Then the sample covariance matrix is

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

Estimation problem and random matrices



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

- The desired estimation $\|\Sigma_n - I\| \le \varepsilon$ is equivalent to saying that $\frac{1}{\sqrt{n}} A$ is an almost isometric embedding $\mathbb{R}^p \to \mathbb{R}^n$:

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

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

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

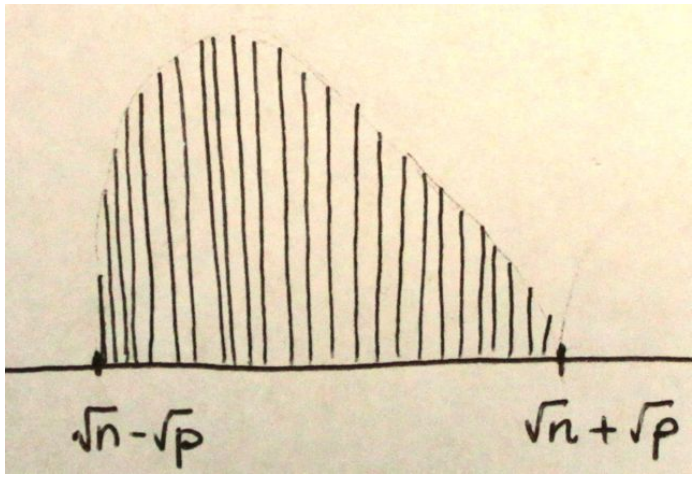
Question. What random matrices with independent rows are almost isometric embeddings?

Random matrices with independent entries

- Simplest example: Gaussian distributions.
  A is a $p \times n$ random matrix with independent $N(0,1)$ entries.
  Σ_n is called Wishart matrix.
- Random matrix theory in the asymptotic regime $n, p \to \infty$:

Theorem (Bai-Yin Law) When $n, p \to \infty$, $n/p \to$ const, one has

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



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 isometric embedding.
- Formally, the sample covariance matrix $\Sigma_n = \frac{1}{n} A^T A$ nicely approximates the actual covariance matrix I:

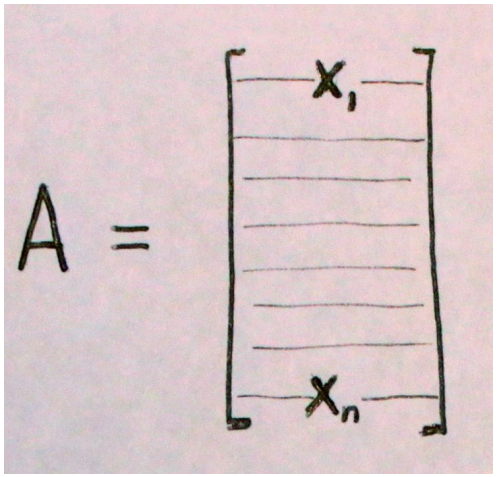
$$\|\Sigma_n - I\| \approx 2\sqrt{\frac{p}{n}} + \frac{p}{n}.$$

Answer to the Estimation Problem for Gaussian distributions: sample size $n(p) \sim p$ suffices to estimate the covariance matrix by a sample covariance matrix.

Random matrices with independent rows

- However, many distributions of interest do not have independent coordinates. Thus the random matrix A has independent rows (samples), but not independent entries in each row.

Problem. Study the spectrum properties of random matrices with independent rows. When do such $n \times p$ matrices A produce almost isometric embeddings?



High dimensional distributions

Under appropriate moment assumptions on the distribution (of the rows), are there results similar to Bai-Yin?

Definition. A distribution of $\mathbf{X}$ in $\mathbb{R}^p$ is subgaussian if all its one-dimensional marginals are subgaussian random variables:

$$\mathbb{P}\{|\langle \mathbf{X}, x\rangle| \geq t\} \leq 2\exp(-ct^2).$$

- Similarly we define subexponential distributions (with tails $2\exp(-ct)$), distributions with finite moments, etc. We thus always assess a distribution by its one-dimensional marginals.
- Examples: The standard normal distribution, the uniform distributions on round ball, cube of unit vol are subgaussian.
- The uniform distribution on any convex body of unit volume is sub-exponential (follows from Brunn-Minkowski inequality, see Borell's lemma). Discrete distributions are usually not even subexponential unless they are supported by exponentially many points.

Random matrices with independent subgaussian rows

Proposition (Random matrices with subgaussian rows). Let A be an $n \times p$ matrix whose rows $\mathbf{X}_k$ are independent sub-gaussian isotropic random vectors in $\mathbb{R}^p$. Then with high probability,

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

- As before, this yields that the sample covariance matrix $\Sigma_n = \frac{1}{n} A^T A$ approximates the actual covariance matrix I:

$$\|\Sigma_n - I\| \le C\sqrt{\frac{p}{n}} + C\frac{p}{n}.$$

- Answer to the Estimation Problem for subgaussian distributions is same as for Gaussian ones: sample size $n(p) \sim p$ suffices to estimate the covariance matrix by a sample covariance matrix.

Random matrices with independent subgaussian rows

Proposition (Random matrices with subgaussian rows). Let A be an $n \times p$ matrix whose rows $\mathbf{X}_k$ are independent sub-gaussian isotropic random vectors in $\mathbb{R}^p$. Then with high probability,

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

Proof (ε-net argument). As we know, the conclusion is equivalent to saying that $\frac{1}{\sqrt{n}}A$ is an almost isometric embedding.

Equivalently, we need to show that $\|Ax\|_2^2$ is close to its expected value n for every unit vector x. But

$$\|Ax\|_2^2 = \sum_{k=1}^{n} \langle \mathbf{X}_k, x \rangle^2$$

is a sum of independent subexponential random variables.

Exponential deviation inequalities (Bernstein's) yield that $\|Ax\|^2 \approx n$ with high probability. Conclude by taking union bound over x in some fixed net of the sphere S^{p-1} and approximation.

Beyond sub-subgaussian

- This argument fails for anything weaker than sub-gaussian distributions – exponential deviation inequalities will fail. Different ideas are needed to address the Estimation Problem for distributions with heavier tails.
- Boundedness assumption: we will assume throughout the rest of this talk that the distribution is supported in a centered ball of radius $O(\sqrt{p})$. Most of the (isotropic) distribution always lies in that ball, as $\mathbb{E}\|\mathbf{X}\|_2^2 = p$.

Random matrices with heavy-tailed rows

Under no moment assumptions at all, we have:

Theorem (Random matrices with heavy tails). Let A be an $n \times p$ matrix whose rows $\mathbf{X}_k$ are independent isotropic random vectors in $\mathbb{R}^p$. Then with high probability,

$$\sqrt{n} - C\sqrt{p \log p} \le s_{\min}(A) \le s_{\max}(A) \le \sqrt{n} + C\sqrt{p \log p}.$$

- $\log p$ is needed (uniform distribution on p orthogonal vectors).
- As before, this yields that the sample covariance matrix $\Sigma_n = \frac{1}{n} A^T A$ approximates the actual covariance matrix I:

$$\|\Sigma_n - I\| \le C\sqrt{\frac{p \log p}{n}} \quad \text{for } n \ge p.$$

  This result was proved by Rudelson'00 (Bourgain'99: $\log^3 p$).
- The answer to the Estimation Problem for heavy-tailed distributions is requires a logarithmic oversampling: a sample size $n(p) \sim p \log p$ suffices to estimate the covariance matrix.

Random matrices with heavy-tailed rows

Theorem (Random matrices with heavy tails). Let A be an $n \times p$ matrix whose rows $\mathbf{X}_k$ are independent isotropic random vectors in $\mathbb{R}^p$. Then with high probability,

$$\sqrt{n} - C\sqrt{p \log p} \le s_{\min}(A) \le s_{\max}(A) \le \sqrt{n} + C\sqrt{p \log p}.$$

Proof There are now several ways to prove this result. The most straightforward argument: Ashlwede-Winter's approach. It directly addresses the Estimation Problem. The sample covariance matrix

$$\Sigma_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{X}_k \mathbf{X}_k^T$$

is a sum of independent random matrices $\mathbf{X}_k \mathbf{X}_k^T$. One can prove and use versions of classical deviation inequalities (Chernoff, Hoeffding, Bernstein, Bennett . . .) for sums of random *matrices*. Proofs are similar – exponentiating and estimating m.g.f. using trace inequalities (Golden-Thompson). See Tropp'10.

When is the logarithmic oversampling needed?

Problem (when is logarithmic oversampling needed?) Classify the distributions in $\mathbb{R}^p$ for which the sample size $n(p) \sim p$ suffices to estimate the covariance matrix by the sample covariance matrix.

- What we know: for general distributions, logarithmic oversampling is needed: $n(p) \sim p \log p$ by Rudelson's theorem. For subgaussian distributions, not needed: $n(p) \sim p$.
- It was recently shown that $n(p) \sim p$ for sub-exponential distributions: Adamczak, Litvak, Pajor, Tomczak'09. This includes uniform distributions on all convex bodies.
- But there is still a big gap between the distributions that do not require the logarithmic oversampling (convex bodies) and those that do require (very discrete).
- How to close this gap? We conjecture that for most distributions, $n(p) \sim p$. For example, this should hold under any non-trivial moment assumptions:

The logarithmic oversampling is almost never needed?

Conjecture. Consider a distribution in $\mathbb{R}^p$ with bounded q-th moment for some $q > 2$, i.e. $\mathbb{E}|\langle \mathbf{X}, x\rangle|^q \le C^q$ for all unit vectors x. Then the sample size $n \sim p$ suffices for estimation of the covariance matrix Σ by the sample covariance matrix Σ_n w.h.p.:

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

- Recall that any isotropic distributions has a bounded second moment. The conjecture says that a slightly higher moment should suffice for estimation without logarithmic oversampling.

The logarithmic oversampling is almost never needed

Theorem (Covariance). Consider a distribution in $\mathbb{R}^p$ with bounded q-th moment for some $q > 4$. Then the sample covariance matrix Σ_n approximates covariance matrix: with high probability,

$$\|\Sigma_n - \Sigma\| \le (\log\log p)^2 \left(\frac{p}{n}\right)^{\frac{1}{2}-\frac{2}{q}}.$$

- As a consequence, the sample size $n \sim (\log\log p)^{C_q} p$ suffices for covariance estimation: $\|\Sigma_n - \Sigma\| \le \varepsilon$.

Estimation of moments of marginals

- Once we know Σ we know the variances of all one-dimensional marginals: $\langle \Sigma x, x\rangle = \langle \mathbb{E}\mathbf{X}\mathbf{X}^T x, x\rangle = \mathbb{E}\langle \mathbf{X}, x\rangle^2$.
- More generally, we can estimate r-th moments of marginals:

Theorem (Marginals). Consider a random vector $\mathbf{X}$ in $\mathbb{R}^p$ with bounded $4r$-th moment. Take a sample of size $n \sim p$ if $r \in [1,2)$ and $n \sim p^{r/2}$ if $r \in (2,\infty)$. Then with high probability,

$$\sup_{x \in S^{p-1}} \left| \frac{1}{n} \sum_{k=1}^{n} |\langle \mathbf{X}, x\rangle|^r - \mathbb{E}|\langle \mathbf{X}, x\rangle|^r \right| \le \varepsilon.$$

- The sample size n has optimal order for all r.
- For subexponential distributions, this result is due to Adamczak, Litvak, Pajor, Tomczak'09. Without extra moment assumptions (except the r-th), a logarithmic oversampling is needed as before. The optimal sample size in this case is $n \sim p^{r/2} \log p$ due to Guedon, Rudelson'07.

Norms of random operators

Corollary (Norms of random operators). Let A be an $n \times p$ matrix whose rows $\mathbf{X}_k$ are independent random vectors in $\mathbb{R}^p$ with bounded $4r$-th moment, $r \geq 2$. Then with high probability,

$$\|A\|_{\ell_2 \to \ell_r} \lesssim n^{1/2} + p^{1/r}.$$

- This result is also optimal. Conjectured to hold for $r = 2$.
- For subexponential distributions, this result is due to Adamczak, Litvak, Pajor, Tomczak'09. Without extra moment assumptions (except the r-th), a logarithmic oversampling is needed as before.

Heuristics of the argument: structure of divergent series

- Two new ingredients in the proofs of these results:
  (1) structure of slowly divergent series;
  (2) a new decoupling technique.
- Consider a simpler problem: for a random vector with heavy tails, we want to show that $\|\Sigma_n\| = O(1)$:

$$\|\Sigma_n\| = \sup_{x \in S^{n-1}} \frac{1}{n} \sum_{k=1}^{n} \langle \mathbf{X}_k, x \rangle^2 = O(1).$$

  This is a stochastic process indexed by vectors $x \in S^{n-1}$.
- For each fixed x, we have to control the sum of independent random variables $\sum_k \langle \mathbf{X}_k, x \rangle^2$. Unfortunately, because of the heavy tails of these random variables, we can only control the sum with a polynomial rather than exponential probability $1 - n^{-O(1)}$. This is too weak for uniform control over x in the sphere S^{p-1} where ε-nets have exponential sizes in p.

Sums of independent heavy-tailed random variables

- This brings us to a basic question in probability theory: control a sum of independent heavy-tailed random variables Z_k.
- Here we follow a simple "combinatorial" approach. Suppose

$$\frac{1}{n}\sum_{k=1}^{n} Z_k \gg 1.$$

  Try and locate some structure in the terms Z_k that is responsible for the largeness of the sum.
- Often one can find an ideal structure: a subset of very large terms Z_k. Namely, suppose there is $I \subset [n]$, $|I| = n_0$ such that

$$Z_k \geq 4\frac{n}{n_0} \quad \text{for } k \in I.$$

  (we can always locate an ideal structure loosing $\log n$).

Sums of independent heavy-tailed random variables

Ideal structure: a subset I, $|I| = n_0$, such that $Z_k \geq 4\frac{n}{n_0}$ for $k \in I$.

- Advantage of the ideal structure: the probability that it exists can be easily bounded. Even if Z_k have just the first moment, say $\mathbb{E}Z_k = 1$:
- By independence, Markov's inequality and union bound over I,

$$\mathbb{P}\{\text{ideal structure exists}\} \leq \binom{n}{n_0}\left(\frac{n_0}{4n}\right)^{n_0} \leq e^{-2n_0}.$$

We get an exponential probability despite the heavy tails.

Combinatorial approach for stochastic processes

- Let us see how the combinatorial approach works for controlling stochastic processes. Assume for some $x \in S^{n-1}$

$$\frac{1}{n}\sum_{k=1}^{n}\langle \mathbf{X}_k, x\rangle^2 \gg 1.$$

- Suppose we can locate an ideal structure responsible for this: a subset I, $|I| = n_0$, such that $\langle \mathbf{X}_k, x\rangle^2 \geq 4\frac{n}{n_0}$ for $k \in I$. As we know,

$$\mathbb{P}\{\text{ideal structure}\} \leq e^{-2n_0}.$$

- This is still not strong enough to take union bound over all x in some net of the sphere S^{p-1} which has cardinality e^n.
- Dimension reduction: By projecting x onto $E = \text{span}(\mathbf{X}_k)_{k\in I}$ we can automatically assume that $x \in E$. This subspace has dimension n_0. Its ε-net has cardinality e^{n_0} which is OK!
- Unfortunately, $x \in E$ becomes random, correlated with $\mathbf{X}_k$'s.
- Decoupling can make x depend on a half of $\mathbf{X}_k$'s (random selection a la Maurey). Condition on this half, finish the proof.

Combinatorial approach for stochastic processes

- This argument yields the optimal Marginal Theorem (on estimation of r-th moments of one-dimensional marginals).
- Generally, in locating the ideal structure one looses a $\log p$ factor. To loose just $\log \log p$ as in the Covariance Theorem, one has to locate a structure that's weaker (thus harder to find) than the ideal structure. This requires a structural theorem for series that diverge slower than the iterated logarithm.

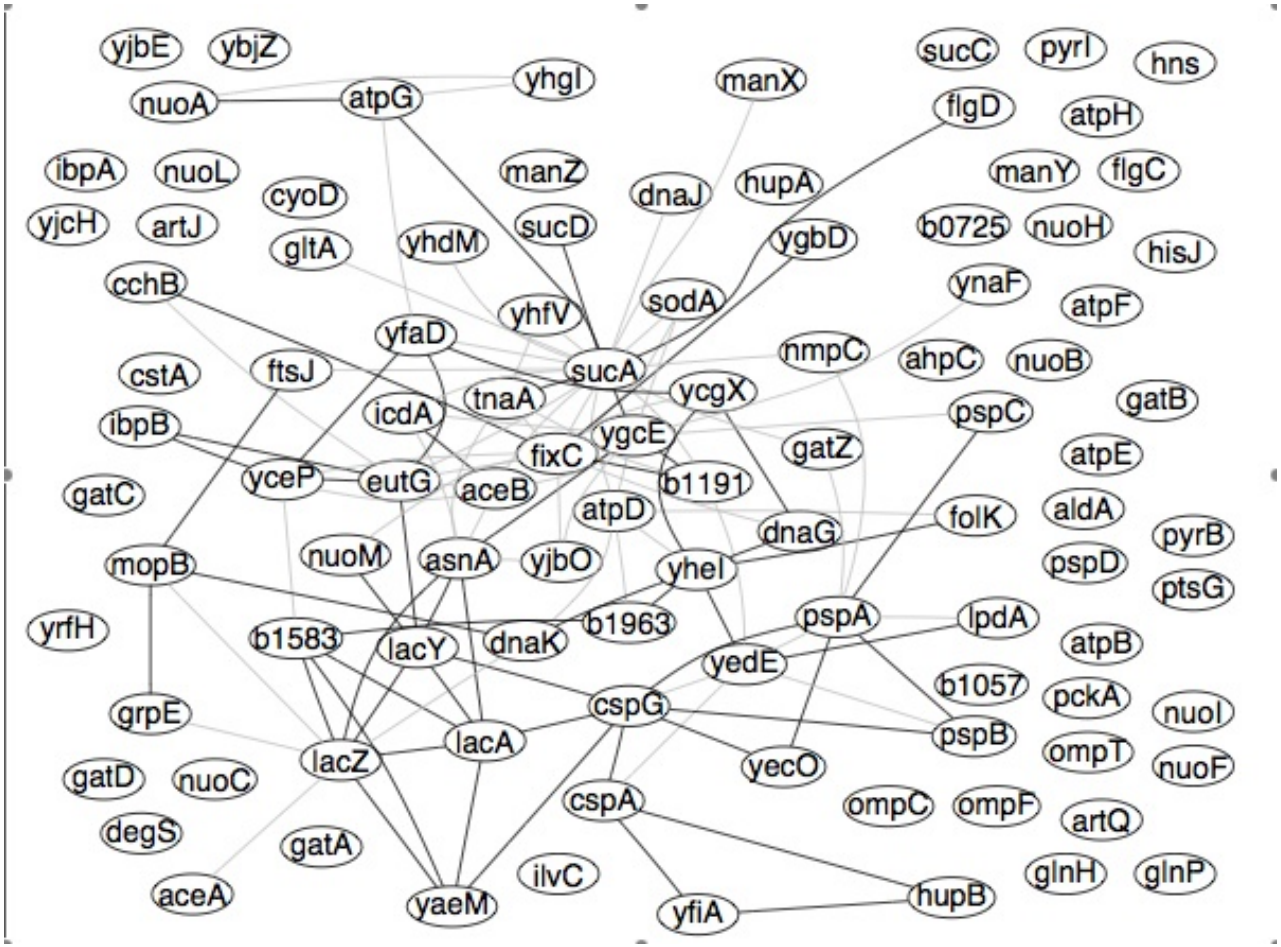
Sparse estimation of covariance matrices

- A variety of practical applications (genomics, pattern recognition, etc.) require very small sample sizes compared with the 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, Σ is often 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 $s \leq p$ 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 problem is nontrivial even for Gaussian distributions, and even if we know the location of the non-zero entries of Σ. Let's assume this (otherwise take the biggest entries of Σ_n).
- Method: compute the sample covariance matrix Σ_n. Zero out all entries that are a priori known to be zero. The resulting sparse matrix $M \cdot \Sigma_n$ 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

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

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

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

  This matches the Theorem in the non-sparse case $s = p$.
- Note the mild, logarithmic dependence on the dimension p and the optimal dependence on the sparsity s.
- A logarithmic factor is needed for $s = 1$, when $M = I$.
- As a consequence, sample size $n \sim s \log^6 p$ suffices for sparse estimation. In the sparse case $s \ll p$, we have $n \ll p$.

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.

Complexity of matrix norm

- Sparse Estimation Theorem would follow by an ε-net argument if the norm of a sparse matrix can be computed on a small set.
- As is well known, the operator norm of an $p \times p$ matrix A can be computed on an $\frac{1}{2}$-net $\mathcal{N}$ of the unit sphere S^{p-1}

$$\|A\| \sim \max_{x \in \mathcal{N}} \|Ax\|_2$$

and one can construct such net with cardinality $|\mathcal{N}| \leq e^{cp}$.
- Can one reduce the size of $\mathcal{N}$ for sparse matrices?

Question (discretizing the norm of sparse matrices). Does there exist a subset $\mathcal{N}$ of S^{p-1} such that, for every $p \times p$ matrix A with at most s nonzero entries in each row and column, one has

$$\|A\| \sim \max_{x \in \mathcal{N}} \|Ax\|_2$$

and with cardinality $|\mathcal{N}| \leq (Cp/s)^s \leq p^{Cs}$?

Gaussian Chaos

- Since we don't know how to answer this question, the proof of the estimation theorem takes a different route – through estimating a Gaussian chaos.
- A gaussian chaos arises naturally when one tries to compute the operator norm of a sample covariance matrix $\Sigma_n = \frac{1}{n}\sum_{k=1}^n \mathbf{X}_k \mathbf{X}_k^T$:

$$\|\Sigma_n\| = \sup_{x \in S^{p-1}} \langle \Sigma_n x, x \rangle = \sum_{i,j=1} \Sigma_n(i,j) x_i x_j = \frac{1}{n} \sum_{k,i,j} X_{ki} X_{kj} x_i x_j$$

where X_{kj} are Gaussian random variables (the coordinates of the sampled points from the Gaussian distribution).

- Argument: (1) decoupling; (2) "combinatorial" approach to estimation, classifying x according to the measure of its sparsity – similar to [Schechtman'04] and many later papers.

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.