Learning Gaussian Covariance Robustly

Daniel M. Kane

Departments of CS/Math
University of California, San Diego
dakane@ucsd.edu

June 23rd, 2019

Outline

- Problem Setup
- Rough Estimates
- Refined Estimates
- Unknown Mean

Basic Problem

- Consider $G = N(0, \Sigma) \subset \mathbb{R}^n$.
- Given N samples, ϵ-fraction adversarially corrupted.
- Learn approximation to Σ.

Error Metric

How closely can we expect to learn Σ?

Error Metric

How closely can we expect to learn Σ?
Can't learn G to better than ϵ total variation.

Error Metric

How closely can we expect to learn Σ?
Can't learn G to better than ϵ total variation.

$$d_{TV}(N(0,\Sigma),N(0,\Sigma')) = \Theta(\min(1,\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F)),$$

where

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2.$$

Error Metric

How closely can we expect to learn Σ?
Can't learn G to better than ϵ total variation.

$$d_{TV}(N(0,\Sigma),N(0,\Sigma')) = \Theta(\min(1,\|\Sigma^{-1/2}\Sigma'\Sigma^{-1/2} - I\|_F)),$$

where

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2.$$

Hope get estimate $\hat{\Sigma}$ so that:

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\|_F = \tilde{O}(\epsilon).$$

Basic Technique

Learning the mean of a Gaussian is equivalent to

- Learning $\mathbb{E}[L(G)]$ for degree-1 polynomials L.
- Learning the first moments of G.

Basic Technique

Learning the mean of a Gaussian is equivalent to

- Learning $\mathbb{E}[L(G)]$ for degree-1 polynomials L.
- Learning the first moments of G.

Learning the covariance of a mean 0 Gaussian is equivalent to:

- Learning $\mathbb{E}[p(G)]$ for even, degree-2 polynomials p.
- Learning the second moments of G.
- Learning $\mathbb{E}[GG^T]$.

Basic Technique

Learning the mean of a Gaussian is equivalent to

- Learning $\mathbb{E}[L(G)]$ for degree-1 polynomials L.
- Learning the first moments of G.

Learning the covariance of a mean 0 Gaussian is equivalent to:

- Learning $\mathbb{E}[p(G)]$ for even, degree-2 polynomials p.
- Learning the second moments of G.
- Learning $\mathbb{E}[GG^T]$.

We will use the last of these formulations.

Robust Mean Estimation

We have reduced the problem to robustly estimating the mean of the n^2-dimensional random variable $X = GG^T$. Since $\mathrm{Cov}(G) = \Sigma = \mathbb{E}[X]$.

Robust Mean Estimation

We have reduced the problem to robustly estimating the mean of the n^2-dimensional random variable $X = GG^T$. Since $\mathrm{Cov}(G) = \Sigma = \mathbb{E}[X]$.

Let $\mathbf{\Sigma} = \mathrm{Cov}(X)$. If $\mathbf{\Sigma} \ll I_{n^2}$, can learn $\mathbb{E}[X]$ to L^2 error (and thus, Σ to Frobenius error) $O(\sqrt{\epsilon})$.

Robust Mean Estimation

We have reduced the problem to robustly estimating the mean of the n^2-dimensional random variable $X = GG^T$. Since $\mathrm{Cov}(G) = \Sigma = \mathbb{E}[X]$.

Let $\mathbf{\Sigma} = \mathrm{Cov}(X)$. If $\mathbf{\Sigma} \ll I_{n^2}$, can learn $\mathbb{E}[X]$ to L^2 error (and thus, Σ to Frobenius error) $O(\sqrt{\epsilon})$.

So, what is $\mathbf{\Sigma}$?

Computing Σ

- Suppose that y_i are an orthonormal basis of linear functions of G.
  - $\mathrm{Cov}(y_i, y_j) = \delta_{i,j}$

Computing Σ

- Suppose that y_i are an orthonormal basis of linear functions of G.
  - $\mathrm{Cov}(y_i, y_j) = \delta_{i,j}$
- $y_i y_j (i \neq j)$ and $(y_i^2 - 1)/\sqrt{2}$ form an orthonormal basis for even degree-2 polynomials of G.

Computing Σ

- Suppose that y_i are an orthonormal basis of linear functions of G.
  - $\mathrm{Cov}(y_i, y_j) = \delta_{i,j}$
- $y_i y_j (i \neq j)$ and $(y_i^2 - 1)/\sqrt{2}$ form an orthonormal basis for even degree-2 polynomials of G.
- For matrix A,

$$A^{flat} \Sigma A^{flat} = \mathrm{Var}(A \cdot X) = \mathrm{Var}(G^T A G)$$
$$= 2 \left\| \Sigma^{1/2} \left(\frac{A + A^T}{2} \right) \Sigma^{1/2} \right\|_F^2 .$$

Computing $\mathbf{\Sigma}$

- Suppose that y_i are an orthonormal basis of linear functions of G.
  - $\mathrm{Cov}(y_i, y_j) = \delta_{i,j}$
- $y_i y_j (i \neq j)$ and $(y_i^2 - 1)/\sqrt{2}$ form an orthonormal basis for even degree-2 polynomials of G.
- For matrix A,

$$A^{flat} \mathbf{\Sigma} A^{flat} = \mathrm{Var}(A \cdot X) = \mathrm{Var}(G^T A G)$$
$$= 2 \left\| \Sigma^{1/2} \left(\frac{A + A^T}{2} \right) \Sigma^{1/2} \right\|_F^2 .$$

So, for example, if $\Sigma \leq I$, $\mathbf{\Sigma} \ll I$.

Bootstrapping

- To learn Σ, need to learn $\mathbb{E}[X]$ robustly.
- Can learn $\mathbb{E}[X]$ robustly, if we have an upper bound on $\boldsymbol{\Sigma}$.
- Can find $\boldsymbol{\Sigma}$ if we know Σ.

Bootstrapping

- To learn Σ, need to learn $\mathbb{E}[X]$ robustly.
- Can learn $\mathbb{E}[X]$ robustly, if we have an upper bound on $\boldsymbol{\Sigma}$.
- Can find $\boldsymbol{\Sigma}$ if we know Σ.

Bootstrap better and better approximations to Σ!

Upper Bounds

Critical Point: If $\Sigma \leq \Sigma_0$, then $\mathbf{\Sigma} \leq \mathbf{\Sigma}_0$, i.e.

$$2\left\|\Sigma^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma^{1/2}\right\|_F^2 \leq 2\left\|\Sigma_0^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma_0^{1/2}\right\|_F^2$$

for all A.

Upper Bounds

Critical Point: If $\Sigma \leq \Sigma_0$, then $\boldsymbol{\Sigma} \leq \boldsymbol{\Sigma}_0$, i.e.

$$2\left\|\Sigma^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma^{1/2}\right\|_F^2 \leq 2\left\|\Sigma_0^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma_0^{1/2}\right\|_F^2$$

for all A.

So if $\Sigma \leq \Sigma_0$, then $\mathrm{Cov}(\Sigma_0^{-1/2} X \Sigma_0^{-1/2}) \ll I_{n^2}$. Can get estimate $\hat{\Sigma}$ with

$$\left\|\Sigma_0^{-1/2}\left(\hat{\Sigma}-\Sigma\right)\Sigma_0^{-1/2}\right\|_F = O(\sqrt{\epsilon}).$$

Upper Bounds

Critical Point: If $\Sigma \leq \Sigma_0$, then $\mathbf{\Sigma} \leq \mathbf{\Sigma}_0$, i.e.

$$2\left\|\Sigma^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma^{1/2}\right\|_F^2 \leq 2\left\|\Sigma_0^{1/2}\left(\frac{A+A^T}{2}\right)\Sigma_0^{1/2}\right\|_F^2$$

for all A.

So if $\Sigma \leq \Sigma_0$, then $\mathrm{Cov}(\Sigma_0^{-1/2} X \Sigma_0^{-1/2}) \ll I_{n^2}$. Can get estimate $\hat{\Sigma}$ with

$$\left\|\Sigma_0^{-1/2}\left(\hat{\Sigma}-\Sigma\right)\Sigma_0^{-1/2}\right\|_F = O(\sqrt{\epsilon}).$$

So $\hat{\Sigma} = \Sigma + O(\sqrt{\epsilon})\Sigma_0$.

Iteration

- Start with some upper bound $\Sigma_0 \geq \Sigma$ (twice the sample covariance works with high probability).
- Get approximation $\hat{\Sigma}_0$.

Iteration

- Start with some upper bound $\Sigma_0 \geq \Sigma$ (twice the sample covariance works with high probability).
- Get approximation $\hat{\Sigma}_0$.
- Use $\Sigma_1 = \hat{\Sigma}_0 + C\sqrt{\epsilon}\Sigma_0$ as new upper bound.
- Get approximation $\hat{\Sigma}_1$.

Iteration

- Start with some upper bound $\Sigma_0 \geq \Sigma$ (twice the sample covariance works with high probability).
- Get approximation $\hat{\Sigma}_0$.
- Use $\Sigma_1 = \hat{\Sigma}_0 + C\sqrt{\epsilon}\Sigma_0$ as new upper bound.
- Get approximation $\hat{\Sigma}_1$.
- Use $\Sigma_2 = \hat{\Sigma}_1 + C\sqrt{\epsilon}\Sigma_1$ as new upper bound.
- ...

Iteration

- Start with some upper bound $\Sigma_0 \geq \Sigma$ (twice the sample covariance works with high probability).
- Get approximation $\hat{\Sigma}_0$.
- Use $\Sigma_1 = \hat{\Sigma}_0 + C\sqrt{\epsilon}\Sigma_0$ as new upper bound.
- Get approximation $\hat{\Sigma}_1$.
- Use $\Sigma_2 = \hat{\Sigma}_1 + C\sqrt{\epsilon}\Sigma_1$ as new upper bound.
- ...

Have $\Sigma_{i+1} \leq \Sigma + O(\sqrt{\epsilon})\Sigma_i$. Eventually get $\Sigma_\infty \leq \Sigma(1 + O(\sqrt{\epsilon}))$, and $\hat{\Sigma}$ with

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I\|_F = O(\sqrt{\epsilon}).$$

Error Idea

- Have error $O(\sqrt{\epsilon})$.
  - Best possible using only bounds on $\mathrm{Cov}(X)$.

Error Idea

- Have error $O(\sqrt{\epsilon})$.
  - Best possible using only bounds on $\mathrm{Cov}(X)$.
- Hope to do better given:
  - Accurate approximation to $\mathrm{Cov}(X)$.
  - Tail bounds for X.

Error Idea

- Have error $O(\sqrt{\epsilon})$.
  - Best possible using only bounds on $\mathrm{Cov}(X)$.
- Hope to do better given:
  - Accurate approximation to $\mathrm{Cov}(X)$.
  - Tail bounds for X.

Simplifying Assumption: $\Sigma \approx I$.

Concentration

Standard Result: If p is a degree-2 polynomial with $\mathrm{Var}(p(G)) = O(1)$, then

$$\Pr(|p(G) - \mathbb{E}[p(G)]| > T) = O(\exp(-\Omega(T))).$$

Concentration

Standard Result: If p is a degree-2 polynomial with $\mathrm{Var}(p(G)) = O(1)$, then

$$\Pr(|p(G) - \mathbb{E}[p(G)]| > T) = O(\exp(-\Omega(T))).$$

Therefore, X has exponential concentration about its mean in any direction.

Setup

- Know Σ to error $O(\delta)$ in Frobenius norm

Setup

- Know Σ to error $O(\delta)$ in Frobenius norm
- Compute $\mathbf{\Sigma}$ to error $O(\delta)$ in operator norm

Setup

- Know Σ to error $O(\delta)$ in Frobenius norm
- Compute $\boldsymbol{\Sigma}$ to error $O(\delta)$ in operator norm
- Compute Σ to error $O(\sqrt{\epsilon\delta} + \epsilon \log(1/\epsilon))$ in Frobenius norm

Setup

- Know Σ to error $O(\delta)$ in Frobenius norm
- Compute $\boldsymbol{\Sigma}$ to error $O(\delta)$ in operator norm
- Compute Σ to error $O(\sqrt{\epsilon\delta} + \epsilon \log(1/\epsilon))$ in Frobenius norm
- Iterate to get error $O(\epsilon \log(1/\epsilon))$

Combine

- Last Talk: If $\Sigma = I$, robustly learn μ.
- This Talk: If $\mu = 0$, robustly learn Σ.

Combine

- Last Talk: If $\Sigma = I$, robustly learn μ.
- This Talk: If $\mu = 0$, robustly learn Σ.
- Question: What if neither Σ nor μ is known?

Trick

- Consider differences of pairs of samples $G_{2i} - G_{2i+1}$.

Trick

- Consider differences of pairs of samples $G_{2i} - G_{2i+1}$.
- Distributed as $N(0, 2\Sigma)$, with 2ϵ error.

Trick

- Consider differences of pairs of samples $G_{2i} - G_{2i+1}$.
- Distributed as $N(0, 2\Sigma)$, with 2ϵ error.
- Use to learn $\hat{\Sigma}$, an approximation to Σ with $O(\epsilon \log(1/\epsilon))$ error.

Trick

- Consider differences of pairs of samples $G_{2i} - G_{2i+1}$.
- Distributed as $N(0, 2\Sigma)$, with 2ϵ error.
- Use to learn $\hat{\Sigma}$, an approximation to Σ with $O(\epsilon \log(1/\epsilon))$ error.
- $\hat{\Sigma}^{-1/2} G \approx N(\hat{\Sigma}^{-1/2}\mu, I)$
  - Treat difference as $O(\epsilon \log(1/\epsilon))$ adversarial error.
  - Use to learn approximation to μ.

Trick

- Consider differences of pairs of samples $G_{2i} - G_{2i+1}$.
- Distributed as $N(0, 2\Sigma)$, with 2ϵ error.
- Use to learn $\hat{\Sigma}$, an approximation to Σ with $O(\epsilon \log(1/\epsilon))$ error.
- $\hat{\Sigma}^{-1/2} G \approx N(\hat{\Sigma}^{-1/2}\mu, I)$
  - Treat difference as $O(\epsilon \log(1/\epsilon))$ adversarial error.
  - Use to learn approximation to μ.

Final result: Learn distribution for G to $\tilde{O}(\epsilon)$ error in total variational distance.

Conclusion

We can learn the mean and covariance of an unknown Gaussian robustly. In order to do so, we need to consider the 2nd and 4th moments of the distribution in question. Later we will look into cases where even higher moments are useful.