Lecture 18: Private covariance estimation in high dimensions

December 3, 2019

1 Introduction

In the last lecture, we gave algorithms for privately estimating the mean of a univariate Gaussian with known variance. This begs the natural question: what about covariance estimation? And what happens in high dimensions? In this lecture, we'll focus on arguably the most interesting of the 4 problems in the "Punnett square", namely, covariance estimation in high dimensions. However, before we do so, we'll briefly cover the other two problems:

1.1 Private mean estimation in high dimensions

Here the question is as follows: give an (ε, δ)-DP algorithm $\mathcal{A}$ which, given n samples $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, I)$ where $\mu \in \mathbb{R}^d$ is unknown, satisfies $\|\mathcal{A}(X_1, \ldots, X_n) - \mu\|_2 \leq \alpha$. Getting the optimal answer with an efficient algorithm with $\delta = 0$ is still unknown, but there is an inefficient algorithm which achieves the right rate [1].

The story is much simpler when we allow for approximate differential privacy. The following algorithm achieves the near optimal rate in terms of d, α, ε: simply run the private univariate estimation algorithm d times with $\alpha' = \alpha/\sqrt{d}$ and privacy parameter $\varepsilon' = \varepsilon/\sqrt{6d \log 1/\delta}$ and $\delta' = \delta/2d$ to estimate each coordinate of the unknown mean. Each individual algorithm is $(\varepsilon/\sqrt{6T \log 1/\delta}, \delta/2d)$-DP. By advanced composition, the overall algorithm is (ε, δ)-DP. Moreover, if the samples are from a Gaussian, then each algorithm with high probability outputs $\widehat{\mu}_i$ so that $|\widehat{\mu}_i - \mu_i| \leq \alpha/\sqrt{d}$, and thus the final output satisfies $\|\mu - \widehat{\mu}_i\|_2 \leq \alpha$ with high probability.[1] By plugging in the bounds from the previous lecture, the overall complexity is

$$n \gtrsim \frac{d}{\alpha^2} + \frac{\sqrt{d} \cdot \log^{3/2} d/\delta}{\varepsilon} + \frac{d\sqrt{\log 1/\delta \cdot \log \frac{d}{\alpha\varepsilon}}}{\alpha\varepsilon} .$$

Up to log factors, the factors in this bound are optimal up to logarithmic factors [2].

1.2 Private covariance estimation in one dimension

Here we consider the following question: give an (ε, δ)-DP algorithm so that given samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, output $\widehat{\sigma}$ so that $e^{-\alpha} \cdot \sigma \leq \widehat{\sigma} \leq e^{\alpha} \cdot \sigma$ (i.e. learn it to multiplicative error). It turns out the story here is quite similar to univariate mean estimation. If $\sigma = \Theta(1)$, then the truncated empirical second moment of the samples (plus Laplace noise) gives optimal rates. It remains to privately learn a rough scale, i.e. learn σ to constant multiplicative error. To do so, we can do another histogram: if we bucket the real line into infinitely many buckets $B_i = [2^i, 2^{i+1}]$ for $i \in \mathbb{Z}$, then it is not hard to show that if B_i is the bucket with the most samples, then 2^i is a constant factor approximation to σ, with high probability. Thus, if we apply a private histogram algorithm to learn this bucket, then run the algorithm for $\sigma = \Theta(1)$, this gives nearly-optimal rates. You'll work through the details of this algorithm in the homework.

[1] The analysis presented in the previous lecture only gives constant success probability, with with additional polylogarithmic factors can be boosted to have high enough probability to union bound over all d runs of the univariate algorithm

2 Private covariance estimation in high dimensions

Now we come to the main question of the lecture: given samples from $\mathcal{N}(0, \Sigma)$, can we learn Σ? Recall from previous lectures that the most natural notion of closeness for this problem is the "Mahalanobis distance", which is a preconditioned Frobenius norm: $\|A\|_\Sigma = \left\|\Sigma^{-1/2} A \Sigma^{-1/2}\right\|_F$. In one dimension, this exactly corresponds to learning σ to multiplicative error. We have the following rate for learning the covariance in this norm without privacy:

Fact 2.1. *Let $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$. Then the empirical second moment $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$ satisfies*

$$\left\|\Sigma - \widehat{\Sigma}\right\|_\Sigma \lesssim \sqrt{\frac{d^2 + \log 1/\tau}{n}} \,,$$

with probability $1 - \tau$. In particular, if $n \gtrsim \frac{d^2 + \log 1/\tau}{\alpha^2}$, then $\left\|\Sigma - \widehat{\Sigma}\right\|_\Sigma < \alpha$ with probability at least $1 - \delta$.

Our goal will be to match this rate as well as we can with a private algorithm.

2.1 The well-conditioned case

We first consider the simplest case, where the unknown covariance is well-conditioned. Formally, suppose that $I \preceq \Sigma \preceq \kappa I$, where $\kappa = \Theta(1)$ is the *condition number* of the matrix Σ. In this case, as has been the case before, the optimal algorithm (up to logarithmic factors) will simply be a truncated estimator with an appropriate amount of noise added. The natural plug-in estimator is simply the empirical second moment, i.e. $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top$. With probability $1 - \tau$, we know that $\left\|X_i X_i^\top\right\|_F = \|X_i\|_2^2 \lesssim \kappa d \log n/\tau$ for all $i \in [n]$. Note that the Frobenius norm of $\frac{1}{n} X_i X_i^\top$ is the ℓ_2 sensitivity of the estimator. Hence if we define the estimator

$$\widetilde{\Sigma}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \cdot \mathbf{1}\left[\|X_i\|_2^2 \leq C \kappa d \log n/\tau\right] \,,$$

then $\mathcal{A}(X_1, \ldots, X_n) \triangleq \widetilde{\Sigma}(X_1, \ldots, X_n) + N(n, \kappa)$ is ρ-zCDP, where $N(n, \kappa)$ is a matrix with entries which are i.i.d. $\mathcal{N}(0, \sigma(n, \kappa)^2)$ for $\sigma(n, \kappa) \triangleq \frac{C \kappa d \log n/\tau}{n \sqrt{2\rho}}$.

It remains to analyze the error of $\mathcal{A}$. With high probability, we know that $\widehat{\Sigma} = \widetilde{\Sigma}$, so we can use Fact 2.1 to bound the error of $\widetilde{\Sigma}$. It suffices to bound the error caused by $N(\kappa)$. Since the entries of $N(n, \kappa)$ are i.i.d. Gaussians, the Frobenius norm of $N(n, \kappa)$ satisfies the following concentration inequality:

$$\|N(n, \kappa)\|_F \lesssim \frac{C \kappa \log n/\tau}{n \sqrt{2\rho}} \cdot d^2 \sqrt{\log 1/\tau}$$

with probability $1 - \tau$. For $\kappa = \Theta(1)$, we have that that $\|A\|_\Sigma$ and $\|A\|_F$ are equal up to constant factors, and hence combining these guarantees, we obtain:

Lemma 2.2. *Let $\mathcal{A}$ be as defined as above. Then $\mathcal{A}$ is ρ-zCDP, and moreover, with probability $1 - \tau$, we have*

$$\|\mathcal{A}(X_1, \ldots, X_n) - \Sigma\|_\Sigma \lesssim \sqrt{\frac{d^2 + \log 1/\tau}{n}} + \frac{d^2 \kappa \log n/\tau \cdot \sqrt{\log 1/\tau}}{n \sqrt{\rho}} \,.$$

In particular, if $\kappa = \Theta(1)$, and if

$$n \gtrsim \frac{d^2 + \log 1/\tau}{\alpha^2} + \frac{d^2 \log n/\tau \cdot \sqrt{\log 1/\tau}}{\alpha \sqrt{\rho}} \,,$$

then $\|\mathcal{A}(X_1, \ldots, X_n) - \Sigma\|_\Sigma \leq \alpha$ with probability $1 - \tau$.

Recall that we should think of $\sqrt{\rho} \approx \varepsilon$ if ε is our privacy loss parameter, so roughly speaking this says that the price of (ε, δ)-DP in this setting is (up to logarithmic factors), of order $O\left(\frac{d^2}{\alpha \varepsilon}\right)$, and this is tight [2]. As in the univariate case, if $\varepsilon = \Omega(\alpha)$, then again privacy comes essentially "for free".

2.2 The general case

We now turn to the most interesting setting, from an algorithmic perspective, namely, when the scale of the covariance is unknown, i.e. when $\kappa = \omega(1)$. If we were to apply Lemma 2.2 in this setting, we would pay a sample complexity which scales linearly with κ.

Recall that in the univariate setting, this was not so much a problem: there is a simple histogram-based approach that allowed us to get a crude approximation to the scale, and then after rescaling, we could simply apply the estimator for when $\kappa = \Theta(1)$. The problem in high dimensions is that it's not so clear how to make such a histogramming approach work. At a high level, the issue is that not is the scale unknown, but the *direction* in which the scale is unknown is also unknown. For instance, consider the family of covariances of the form $I + (\kappa - 1)uu^\top$ for an arbitrary unit vector $u \in \mathbb{R}^d$. When κ is large, we cannot simply scale down all directions uniformly: we can only afford to scale down the direction $uu^\top$. But it is not clear how to do so without first learning u.

The key idea will be a technique called *recursive private preconditioning*. At a high level, the idea is that the algorithm $\mathcal{A}$ presented above still gives us some weak power, even when κ is large, and we wish to avoid a linear dependence on κ. Specifically, even with relatively few samples, it will give us a rough approximation of any direction which has eigenvalue which is at least $\kappa/10$. We can then precondition away those directions. and slightly decrease the eigenvalues in these directions, so that now the maximum eigenvalue is at most $9\kappa/10$. The effect of this preconditioning is now that if we repeat this process, we can afford to add slightly less noise, as κ has decreased. As a result, we can detect and remove slightly smaller eigenvalues, and iterate this process. Eventually, after $O(\log \kappa)$ iterations, the resulting covariance will have $\kappa = \Theta(1)$, and now finally we can afford to run the naive algorithm.

Let us make this bound formal. Our goal will be to construct a ρ-zCDP algorithm which, given samples from $\mathcal{N}(0, \Sigma)$ for some Σ satisfying $I \preceq \Sigma \preceq \log \kappa$, outputs some symmetric preconditioning matrix P so that

$$I \preceq P\Sigma P \preceq 1000 I \; . \qquad (1)$$

Afterwards, given fresh samples $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$, one can construct the samples $Y_i = PX_i$, so that the Y_i are Gaussian with covariance $P\Sigma P$, and so we can apply the algorithm from Lemma 2.2 to privately output some M so that $\|P\Sigma P - M\|_{P\Sigma P} \leq \alpha$. But then since

$$\begin{aligned}
\|P\Sigma P - M\|_{P\Sigma P}^2 &= \left\|(P\Sigma P)^{-1/2}(P\Sigma P - M)(P\Sigma P)^{-1/2}\right\|_F^2 \\
&= \left\langle (P\Sigma P)^{-1/2}(P\Sigma P - M)(P\Sigma P)^{-1/2}, (P\Sigma P)^{-1/2}(P\Sigma P - M)(P\Sigma P)^{-1/2} \right\rangle \\
&= \operatorname{tr}\left((P\Sigma P)^{-1}(P\Sigma P - M)(P\Sigma P)^{-1}(P\Sigma P - M)\right) \\
&= \left\langle \Sigma^{-1/2}(\Sigma - P^{-1}MP^{-1})\Sigma^{-1/2}, \Sigma^{-1/2}(\Sigma - P^{-1}MP^{-1})\Sigma^{-1/2} \right\rangle \\
&= \left\|\Sigma - P^{-1}MP^{-1}\right\|_\Sigma^2 \; ,
\end{aligned}$$

we conclude that $P^{-1}MP^{-1}$ is a good estimator of Σ in the original Mahalanobis norm.

As an intermediate step to (1), we will give a method which, under the same assumptions, outputs a P so that

$$I \preceq P\Sigma P \preceq \frac{9\kappa}{10} I \; . \qquad (2)$$

If we have such a primitive, then by applying this method $T = O(\log \kappa)$ times, we obtain a sequence of preconditioning matrices $P_1, \ldots, P_T$ and applying composition of privacy, will yield an algorithm that achieves (1) with $P = P_T P_{T-1} \ldots P_1$, and has only a mild dependence on κ.

The key point will be that the algorithm $\mathcal{A}$ described above in fact already allows us to detect large eigenvalues of Σ. This is because it not only gives us good bounds in terms of Frobenius norm, but it also gives us good bounds for spectral norm. These bounds are in fact better, since spectral norm is a less tight measure of distance than Frobenius norm.

We claim that this will allow the following procedure to (essentially) achieve (2): let M be the output of the algorithm $\mathcal{A}$, and let S be the subspace spanned by the eigenvectors of M with eigenvalue at least $\kappa/2$. Let Π be the projection onto this subspace, and let $\Pi_\perp$ be the projection onto the orthogonal subspace, and let $P = \Pi_\perp + \frac{1}{2}\Pi$, that is, we downweight the directions in S, and keep the contribution of the other directions unchanged. Observe that since P is created by postprocessing the result of running a ρ-zCDP algorithm, P is also ρ-zCDP, so we only have prove correctness. Our main claim will be the following:

Theorem 2.3. *Let P be as above, suppose that $\kappa > 1000$, and let n be so that $n \gtrsim d + \log 1/\tau$, and so that $\sigma(n,k)\sqrt{d}\log 1/\tau \lesssim \kappa/100$. Then, with probability at least $1-\tau$, we have that P satisfies $0.99I \preceq P\Sigma P \preceq \frac{4\kappa}{5}I$. In particular, if we let $P' = 1.05P$, we obtain that $I \preceq P'\Sigma P' \preceq \frac{9\kappa}{10}I$.*

Before we prove Theorem 2.3, we will require the following two lemmata. The first states that the empirical covariance converges quickly in spectral norm:

Lemma 2.4. *Let $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$. Then, with probability at least $1-\tau$, we have:*

$$\left\|\Sigma^{-1/2}\left(\Sigma - \widehat{\Sigma}\right)\Sigma^{-1/2}\right\|_{\mathrm{op}} \lesssim \sqrt{\frac{d + \log 1/\tau}{n}}\ .$$

Proof. We've already proven this lemma when $\Sigma = I$ (see Fact 1.2 in Lecture 6), and the general case simply follows because if $Y \sim \mathcal{N}(0,\Sigma)$, then if $X = \Sigma^{-1/2}Y$, then $X \sim \mathcal{N}(0, I)$. □

The second lemma, which is a standard fact in random matrix theory, controls the largest eigenvalue of $N(n,\kappa)$:

Lemma 2.5 (see e.g. [3])**.** *Let d be sufficiently large. Then there exists a universal constant $A > 0$ so that for all $t > A$, we have*

$$\Pr\left[\|N(n,\kappa)\|_{\mathrm{op}} > t\sigma(n,\kappa)\sqrt{d}\right] \leq A\exp(-\Omega(td))\ .$$

Proof of Theorem 2.3. We'll prove the upper bound in Theorem 2.3; the lower bound is similar but slightly more involved so we leave it to the reader (or see [2] for more details). To this end, suppose that u is a unit vector so that $u^\top P\Sigma Pu \geq 4\kappa/5$. We claim that, with high probability, this would imply that $u^\top PMPu > \kappa/2$, which is a contradiction, as PMP has no eigenvalues larger than $\kappa/2$.

We condition on three events all holding simultaneously:

1. we have $\left\|\Sigma^{-1/2}(\Sigma - \widehat{\Sigma})\Sigma^{-1/2}\right\|_{\mathrm{op}} \leq 1/10$,
2. we have $\widehat{\Sigma} = \widetilde{\Sigma}$, and
3. we have $\|N(n,\kappa)\|_2 \leq \kappa/100$.

By setting the internal τ in each part to $\tau/3$ and paying a slight constant overhead in the sample complexity, we may assume that each event holds with probability at least $1-\tau/3$ and so all events hold together with probability at least $1-\tau$.

We first claim that condition (1) implies that $\left\|\Sigma - \widehat{\Sigma}\right\|_{\mathrm{op}} \leq \kappa/10$; this follows as $\Sigma^{-1/2} \succeq \kappa^{-1/2}I$. As a result, since $\|Pu\|_2 \leq 1$, this implies that $\left\|u^\top P\Sigma Pu - u^\top P\widehat{\Sigma}Pu\right\|_2 \leq \kappa/10$, and hence $\left\|u^\top P\Sigma Pu\right\|_2 \geq 3\kappa/4$. As a result, we have

$$\left\|u^\top PMPu\right\|_2 \geq \left\|u^\top P\widetilde{\Sigma}Pu\right\|_2 - \left\|u^\top N(n,\kappa)u\right\|_2 \overset{(a)}{=} \left\|u^\top P\widehat{\Sigma}Pu\right\|_2 - \left\|u^\top N(n,\kappa)u\right\|_2 \overset{(b)}{\geq} \frac{\kappa}{2}\ ,$$

where (a) uses condition (2) and (b) uses condition (3). This completes the proof. □

Let's now quantify the bounds for Theorem 2.3 more explicitly, and get the final bounds for a private algorithm which achieves (1). To satisfy the conditions of Theorem 2.3, it suffices to take

$$n = \widetilde{\Omega}\left(d + \log 1/\tau + \frac{d^{3/2}\log^2(1/\tau)}{\sqrt{\rho}}\right) .$$

We will also need to run the algorithm in Theorem 2.3 for $O(\log \kappa)$ iterations; thus to preserve privacy amongst all iterations and invoke composition, we will need to set the ρ' in every iteration to be $\rho' = \Theta(\rho/\log \kappa)$. Naively, we would also have to use fresh samples in every iteration, and so our sample complexity would also suffer an additional multiplicative $O(\log \kappa)$ factor, however, as [2] shows, if one is slightly careful about analyzing the algorithm, one can show that it is okay to reuse samples across iterations. Similarly, one would naively need to set $\tau' = \Theta(\tau/\log \kappa)$, but again this is avoidable by a more careful analysis. Thus, overall, this yields:

Corollary 2.6 (see [2])**.** *Let $\rho, \tau > 0$, and let $\kappa > 0$. Let $I \preceq \Sigma \preceq \kappa I$. Then, there is a polynomial time ρ-zCDP algorithm which, if given $X_1, \ldots, X_n \sim \mathcal{N}(0, \Sigma)$, and if*

$$n = \widetilde{\Omega}\left(d + \log 1/\tau + \frac{d^{3/2}\log^2(1/\tau)\sqrt{\log \kappa}}{\sqrt{\rho}}\right) ,$$

outputs P satisfying (1) with probability at least $1 - \tau$. Together with Lemma 2.2, this implies that there is a poly-time ρ-zCDP algorithm which, if

$$n \gtrsim \widetilde{\Omega}\left(\frac{d^2 + \log 1/\tau}{\alpha^2} + \frac{d^2 \log n/\tau \cdot \sqrt{\log 1/\tau}}{\alpha\sqrt{\rho}} + \frac{d^{3/2}\log^2(1/\tau)\sqrt{\log \kappa}}{\sqrt{\rho}}\right) ,$$

outputs M so that $\|\Sigma - M\|_\Sigma < \alpha$, with probability at least $1 - \tau$.

Again, in many regimes, the price of privacy (i.e. the last two terms) are dominated by the first, and so privacy comes with little overhead. However, the recursive private preconditioning introduces this (albeit weak) dependence on κ. One can show that such a dependence is necessary for pure DP and zCDP, however it is a great open question whether or not it is necessary for approximate DP.

References

[1] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. *arXiv preprint arXiv:1905.13229*, 2019.

[2] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*, 2018.

[3] T. Tao. *Topics in Random Matrix Theory.* Graduate studies in mathematics. American Mathematical Soc.