Statistics of Big Data, Fall 2021-22

Class notes 11

Review of PCA inference following Boaz's lecture

Dimensionality reduction: If we agree that estimation in high dimension is a tough problem (for supervised and even more for unsupervised), then it seems reasonable to search for low dimensional structure in our data. This is especially relevant if we are willing to assume that such structure exists and captures most/all the relevant information. This is like the sparsity assumption, now for unsupervised learning. The linear version of this searches for "interesting directions" in $\mathbb{R}^p$ where most of the spread of the data is. The classical version is Principal Component Analysis (PCA):

$$\begin{aligned} \hat{v}_1 &= \arg\max_{\|v\|=1} \mathbb{X}v \\ \hat{v}_2 &= \arg\max_{\|v\|=1, v\perp\hat{v}_1} \mathbb{X}v \\ &\vdots \end{aligned}$$

The well known solution is that $\hat{v}_1$ is the eigenvector of $\mathbb{X}^T\mathbb{X}$ with the highest eigenvalue (equivalently the right singular vector of $\mathbb{X} = UDV^T$ with the highest singular value).

As statisticians we want to go beyond calculating things like the principal components to ask statistical questions:

1. How are the top PCs and their eigenvalue distributed in different settings?
2. How can we know whether the top PCs capture "real" structure or noise in the data?
3. How do the answers depend on properties of the problem, especially n, p and their ratio?

The main part of Boaz's lecture discussed the spiked covariance model:

$$x_i = \sum_{j=1}^{L} \sqrt{\lambda_j} z_{ij} v_j + u_i \ , \ \ x_i, v_j, u_i \in \mathbb{R}^p \ , \ \ u_i \sim N(0, \sigma^2 I) \ , \ z_{ij} \sim N(0,1) \ \forall i,j \ , \text{ all independent.}$$

where:

- $v_1, \ldots, v_L$ are the orthonormal "signal" directions in the space
- λ_j is the strength (variance) of signal j
- u_i is the "white" noise that goes in all directions

- z_{ij} is the (relative) signal size for observation i in direction j

In this setting we can write the overall distribution of each observation:

$$\mathbb{R}^p \ni x \sim N(0, \sigma^2 I + \sum_j \lambda_j v_j v_j^T).$$

Now given a data matrix X of n observations drawn from this model, we calculate its top k PCA directions and $\hat{v_1}, \ldots, \hat{v_k}$ and corresponding variance explained/eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_k$, and ask what is the relationship between the empirical and theoretical quantities, and what kind of statistical inference can be done, in what settings.

The main specific questions Boaz discussed in this model:

1. In the case $L = 0$ of no signal, what does the distribution of the eigenvalues look like, and can we define this as a "null" distribution for testing $H_0 : L = 0$?
   The answer, as we saw is the Marchenko-Pastur distribution: For $n, p \to \infty,\ p/n \to \gamma$, we have that the distribution of the eigenvalues converges to the M-P distribution with density:
   $$f(l) = \frac{1}{2\pi\gamma l}\sqrt{b-l}\sqrt{l-a}\ ;\ l \in [a,b]\ ,\ a = (1-\sqrt{\gamma})^2\ ,\ b = (1+\sqrt{\gamma})^2.$$
   Note that M-P has a finite support. This does not quite give us a hypothesis test for finite data (there are more advanced results that allow that), but it gives us a clear understanding of what to expect under the null. As Boaz showed, even for relatively small data sizes the fit to M-P is already very good.

2. In the case $K > 0$, what should be λ_j so we can find them? What about the corresponding v_j? For this we have the very interesting and concrete *phase transition* results:

   - In the same asymptotic regime as M-P, we can expect to find any signal of magnitude $\lambda_j > \sigma^2\sqrt{\gamma}$ in the sense that the empirical eigenvalue $\hat{\lambda}_j$ will "stick out" of the M-P distribution, while for signal of magnitude $\lambda_j < \sigma^2\sqrt{\gamma}$ with high probability $\hat{\lambda}_j$ will be within the M-P range so non-identifiable by definition.
   - The same bound $\lambda > \sigma^2\sqrt{\gamma}$ also guarantees that the directions we find will be "related" to the true directions. This is easiest for $L = 1$ where the result says that in this case
     $$c\hat{o}r^2(v, \hat{v}) \approx \frac{\frac{\lambda^2}{\gamma\sigma^4} - 1}{\frac{\lambda^2}{\gamma\sigma^4} + \frac{\lambda}{\sigma^2}} > 0,$$
     which increases to 1 as λ increases, and decreases to 0 as $\lambda \searrow \sigma^2\sqrt{\gamma}$.

The last part of Boaz's lecture was devoted to learning PCA's and estimating covariance matrices with sparsity:

- In the case the vectors v_j themselves are sparse, he discussed the method of Johnstone and Lu (2009) of simply keeping the high-variance directions and calculating PCs in the reduced space. This provably (under sparsity and other assumptions) allows more efficient learning of PC's if the true vectors are sparse.

- For estimating the covariance in the case the true covariance is sparse, he discussed the simple method of Bickel and Levina (2008) of thresholding the empirical covariance matrix. This also provably shows that the difficult task of covariance estimation can be (somewhat) solved under sparsity assumptions.

The Quality preserving database

Reminder from two weeks ago

Family-wise error rate $FWER = \mathbb{P}(V > 0)$ **(closely related:** $\mathbb{E}(V)$ **)** is the probability of making even one false discovery in the entire corpus. FWER is controlled at level α by the following simple idea: divide α into K pieces $\alpha_1, \ldots, \alpha_K$ such that $\sum_k \alpha_k \leq \alpha$, then test H_{0k} at level α_k. Then it is trivial to see

$$FWER \leq \mathbb{E}(V) = \sum_k \alpha_k \leq \alpha.$$

False Discovery rate (FDR) $\mathbb{E}(\frac{V}{R})$ is the expected proportion of discoveries that are false. The classical approach to control it is the BH approach.

We discussed *publication bias* and how it leads to *Most published research is wrong.*

Solutions we discussed:

- Publishing all results, including negative (then we can correct ourselves for all performed research)
- Encouraging collaboration and data sharing in a way that makes it beneficial to work together in a joint data resource that controls FWER (or some other measures) for all research in the community. Such an idea can use α-spending to control FWER: create an infinite series $\alpha_1, \alpha_2, \ldots$ such that $\sum_j \alpha_j \leq \alpha$, and test hypothesis j at level α_j.

0.1 The QPD schema (Aharoni et al. 2011), (Rosset et al. 2014)

This is a schema to build and maintain a database for scientific research, assuring:

- Statistical validity of all the results being generated on the database
- Usefulness (=maintaining power) for later users

The main idea is to increase the data size as the database is used. In this way we can maintain power even if the levels decrease in α spending!

Assume at time t we have data of size n_{t-1} and remaining α pool of $\alpha \cdot q^{n_{t-1}}$ for given some fixed parameter $q < 1$, say $q = 0.999$. Then the tth test arrives, meaning a scientist has in mind:

- A pair of hypotheses about some parameter θ:

  $$H_{0t} : \theta = \theta_0 \ , \ H_{At} : \theta = \theta_A.$$

  Note this encodes what the test statistic is (through Neyman-Pearson or monotone likelihood), and also what the effect size they think they will find is through θ_A.

- A desired power π_t

At this point we find c_t such that:

$$\alpha_t = \alpha \cdot q^{n_{t-1}}(1 - q^{c_t}),$$

is an appropriate level for getting power π_t for the desired test.

Then after we perform the test and get c_t more samples we have $n_t = n_{t-1} + c_t$ samples and remaining pool of:

$$\alpha \cdot q^{n_{t-1}} - \alpha \cdot q^{n_{t-1}}(1 - q^{c_t}) = \alpha \cdot q^{n_{t-1}+c_t}.$$

Hence by definition our pool will never run out.

Example of power calculation with n_t samples and required effect size θ_t, in the simple normal means case, with known variance σ^2:

$$\begin{aligned}\pi(\alpha_t) &= \mathbb{P}_{\theta_t}\left(\bar{X} \geq Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}}\right) = \mathbb{P}_{\theta_t}\left(\frac{\bar{X} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}} \geq \frac{Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}}\right) = \\ &= 1 - \Phi\left(\frac{Z_{1-\alpha_t}\frac{\sigma}{\sqrt{n_t}} - \theta_t}{\frac{\sigma}{\sqrt{n_t}}}\right)\end{aligned}$$

Theorem (Aharoni et al. 2011):
For many families of testing problems, including:

1. Any string of simple tests that use Neyman-Pearson
2. Tests of normal means

and many others, the simple recipe above guarantees that $c_t \leq c_0$ is bounded in the following sense: A test of a specific effect size $\tilde{\theta}$ at a specific required power $\tilde{\pi}$ will never cost more than $c_0(\tilde{\theta}, \tilde{\pi})$ samples at any time t.

In practice, this leads to diminishing and not only bounded costs (this can in fact be proven rigorously based on the same ideas from the original proof).

Important conclusion: If you come later, you will gain power and/or money.