Colin Adams
Math189R SU20
Homework 5
June 2020

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

1 (Murphy 12.5 - Deriving the Residual Error for PCA) It may be helpful to reference section 12.2.2 of Murphy.

(a) Prove that

$$\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 = \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j.$$

Hint: first consider the case when $k = 2$. Use the fact that $\mathbf{v}_i^\top \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. Recall that $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$.

(b) Now show that

$$J_k = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j.$$

Hint: recall that $\mathbf{v}_j^\top \boldsymbol{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

(c) If $k = d$ there is no truncation, so $J_d = 0$. Use this to show that the error from only using $k < d$ terms is given by

$$J_k = \sum_{j=k+1}^{d} \lambda_j.$$

Hint: partition the sum $\sum_{j=1}^{d} \lambda_j$ into $\sum_{j=1}^{k} \lambda_j$ and $\sum_{j=k+1}^{d} \lambda_j$.

Lots of cut and dry math coming below, so be prepared.

(a) We note the definition of the norm

$$\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 = \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^T \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)$$

and so, if we expand the definition we get,

$$\begin{aligned}
\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 &= \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right)^T \left(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right) \\
&= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{i=1}^{k} z_{ij}\mathbf{x}_i^T\mathbf{v}_j + \sum_{lm}^{k} z_{il}z_{im}\mathbf{v}_l^T\mathbf{v}_m \\
&= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{i=1}^{k} z_{ij}\mathbf{x}_i^T\mathbf{v}_j + \sum_{lm}^{k} z_{il}z_{im}\delta_{lm} && \text{(where } \delta_{ij} \text{ is Kronecker Delta)} \\
&= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{i=1}^{k} z_{ij}\mathbf{x}_i^T\mathbf{v}_j + \sum_{j}^{k} z_{ij}z_{ij} \\
&= \mathbf{x}_i^T\mathbf{x}_i - \sum_{i=1}^{k} \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j && \text{(using def. of } z_{ij}\text{)}
\end{aligned}$$

as desired.

(b) We have

$$\begin{aligned}
J_k &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top\boldsymbol{\Sigma}\mathbf{v}_j && \text{(using def. of } \boldsymbol{\Sigma}\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k}\lambda_j && \text{(since } \boldsymbol{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j \text{ and } \mathbf{v}_j^T\mathbf{v}_j = 1\text{)}
\end{aligned}$$

as desired.

(c) Since $J_d = 0$, from the definition of J_k, we see that

$$0 = J_d = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{d}\lambda_j \quad \text{which imples that} \quad \sum_{j=1}^{d}\lambda_j = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i.$$

Using this fact in the more general relation for J_k, we have

$$\begin{aligned}
J_k &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j \\
&= \sum_{j=1}^{d} \lambda_j - \sum_{j=1}^{k} \lambda_j && \text{(from above)} \\
&= \sum_{j=1}^{k} \lambda_j + \sum_{j=k+1}^{d} \lambda_j - \sum_{j=1}^{k} \lambda_j \\
&= \sum_{j=k+1}^{d} \lambda_j - \sum_{j=1}^{k}
\end{aligned}$$

which tells us the residual error comes from not keeping everything, which isn't super surprising but it is rather cute.

■

2 **(ℓ_1-Regularization)** Consider the ℓ_1 norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|.$$

Draw the norm-ball $B_k = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq k\}$ for $k = 1$. On the same graph, draw the Euclidean norm-ball $A_k = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq k\}$ for $k = 1$ behind the first plot. (Do not need to write any code, draw the graph by hand).

Show that the optimization problem

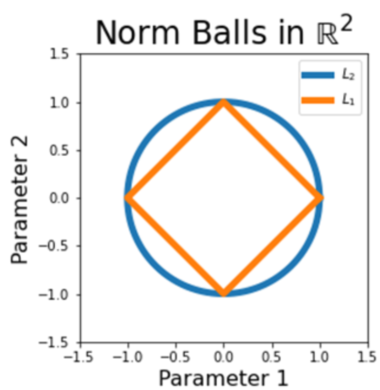
$$\begin{aligned} \text{minimize: } & f(\mathbf{x}) \\ \text{subj. to: } & \|\mathbf{x}\|_p \leq k \end{aligned}$$

is equivalent to

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$$

(hint: create the Lagrangian). With this knowledge, and the plots given above, argue why using ℓ_1 regularization (adding a $\lambda\|\mathbf{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2 regularization for suitably large λ.

The boundary of the norm balls in two dimensions can be visualized as



We have a function $f(\mathbf{x})$ that is constrained to $\|\mathbf{x}\|_p \leq k$ that we are hoping to minimize. We note that the constraint tells us

$$\|\mathbf{x}\|_p - k \leq 0 \qquad \Rightarrow \qquad \lambda\left(\|\mathbf{x}\|_p - k\right) = 0$$

for some λ. Our goal is to find

$$\begin{aligned} \arg\min_{\mathbf{x}} f(\mathbf{x}) &= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \lambda\left(\|\mathbf{x}\|_p - k\right) \\ &= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p \qquad \text{(since } -\lambda k \text{ only shifts function up or down)} \end{aligned}$$

and hence minimizing $f(\mathbf{x})$ is equivalent to minimizing $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$. The L_1 norm gives sparser solutions because there are an infinite number of solutions in which one of the parameters can be zero whereas there is exactly one in the L_2 case. ∎

Extra Credit (Lasso) Show that placing an equal zero-mean Laplace prior on each element of the weights $\boldsymbol{\theta}$ of a model is equivelent to ℓ_1 regularization in the Maximum-a-Posteriori estimate

$$\text{maximize: } p(\boldsymbol{\theta}|\mathscr{D}) = \frac{p(\mathscr{D}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathscr{D})}.$$

Note the form of the Laplace distribution is

$$\text{Lap}(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

where μ is the location parameter and $b > 0$ controls the variance. Draw (by hand) and compare the density $\text{Lap}(x|0,1)$ and the standard normal $\mathcal{N}(x|0,1)$ and suggest why this would lead to sparser solutions than a Gaussian prior on each elements of the weights (which correspond to ℓ_2 regularization).

If we are hoping to maximize $p(\boldsymbol{\theta}|\mathscr{D})$ then this is the same as asking to maximize the $\log p(\boldsymbol{\theta}|\mathscr{D})$. And so

$$\text{maximize: } \log p(\boldsymbol{\theta}|\mathscr{D}) = \log\frac{p(\mathscr{D}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})}{p(\mathscr{D})} = \log p(\mathscr{D}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathscr{D})$$

We ignore the constant term on the right. If we assume that $p(\boldsymbol{\theta}_i) \sim \text{Lap}(\boldsymbol{\theta}_i|\mu = 0, b)$, then

$$\begin{aligned}
\log p(\boldsymbol{\theta}) \propto \log\prod_i \text{Lap}(\boldsymbol{\theta}_i|\mu=0,b) &= \sum_i \log\text{Lap}(\boldsymbol{\theta}_i|\mu=0,b)\\
&= \sum_i \log\frac{1}{2b}\exp\left(-\frac{|\boldsymbol{\theta}_i|}{b}\right)\\
&= -\sum_i \log 2b - \sum_i \frac{|\boldsymbol{\theta}_i|}{b}\\
&= -\sum_i \log 2b - \frac{1}{b}\|\boldsymbol{\theta}\|_1
\end{aligned}$$

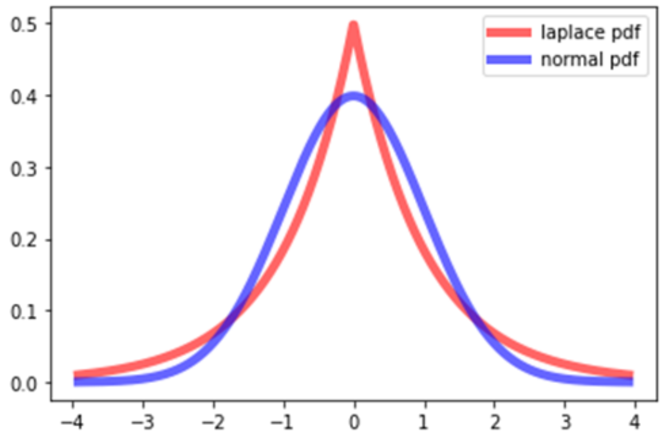
This tells us that—if we ignore any constant terms—that maximizing $p(\boldsymbol{\theta}|\mathscr{D})$ is the same as

$$\text{maximize: } \log p(\boldsymbol{\theta}|\mathscr{D}) = \log p(\mathscr{D}|\boldsymbol{\theta}) - \lambda\|\boldsymbol{\theta}\|_1 \quad \text{where} \quad \lambda \equiv 1/b,$$

or, equivalently,

$$\text{minimize: } \log p(\boldsymbol{\theta}|\mathscr{D}) = -\log p(\mathscr{D}|\boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1.$$

Here is the plot comparing the Laplacian and Gaussian distributions (I'm sorry about the resolution):



This would lead to sparser solutions presumably because more of the distribution is centered around zero than the Gaussian. ∎