Aimee Co
MA189 SP24
Homework 5
2/26/2024

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 $\mathbf{v}_j^\top \mathbf{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

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

(a)

$$\begin{aligned}
\left\| x_i - \sum_{j=1}^{k} z_{ij} v_j \right\|_2^2 &= \left(x_i - \sum_{j=1}^{k} z_{ij} v_j \right)^\top \left(x_i - \sum_{j=1}^{k} z_{ij} v_j \right) \\
&= x_i^\top x_i - \sum_{j=1}^{k} z_{ij} v_j^\top x_i - x_i^\top \sum_{j=1}^{k} z_{ij} v_j + \left(\sum_{j=1}^{k} z_{ij} v_j \right)^\top \left(\sum_{j=1}^{k} z_{ij} v_j \right) \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \left(\sum_{j=1}^{k} z_{ij} v_j \right)^\top \left(\sum_{j=1}^{k} z_{ij} v_j \right) \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j + \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j \\
&= x_i^\top x_i - \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j
\end{aligned}$$

(b)

$$\begin{aligned}
J_k &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^\top x_i - \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{k} v_j^\top \frac{1}{n} \left(\sum_{i=1}^{n} x_i x_i^\top \right) v_j \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{k} v_j^\top \Sigma v_j \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{k} \lambda_j
\end{aligned}$$

(c) Since $J_d = 0$, $\sum_{j=1}^{d} \lambda_j = \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i$. Thus,

$$\begin{aligned}
J_k &= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{d} \lambda_j + \sum_{j=k+1}^{d} \lambda_j \\
&= \sum_{j=k+1}^{d} \lambda_j
\end{aligned}$$

■

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).

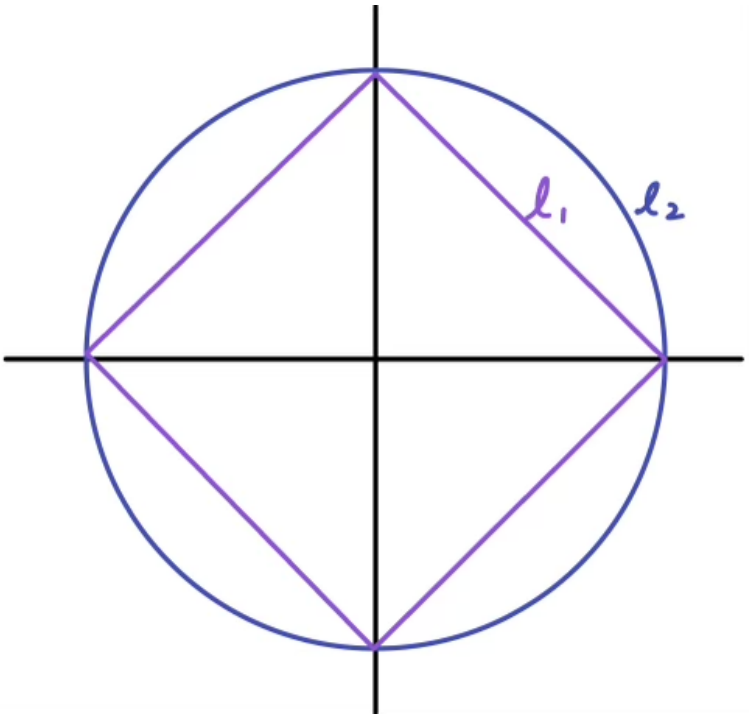
Show that the optimization problem

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

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.



Given the constrained optimization problem

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

we can form its Lagrangian as

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k),$$

where $\lambda \geq 0$ is a Lagrange multiplier. The equivalence to the unconstrained problem

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

is established by optimizing $L(\mathbf{x}, \lambda)$ under the condition that λ enforces the constraint $\|\mathbf{x}\|_p \leq k$.

ℓ_1 regularization, represented by adding a $\lambda\|\mathbf{x}\|_1$ term, promotes sparsity in solutions due to its linear penalty on coefficients. This can lead to coefficients being exactly zero, enhancing model interpretability and reducing complexity. In contrast, ℓ_2 regularization, indicated by a $\lambda\|\mathbf{x}\|_2$ term, tends to shrink coefficients evenly but does not necessarily set them to zero, resulting in less sparse solutions. The geometrical property of the ℓ_1 norm, having corners at zero, is crucial for its sparsity-inducing effect, unlike the smooth contour of the ℓ_2 norm. ■

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

$$\text{maximize: } \mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})}.$$

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).

Given the Maximum-a-Posteriori (MAP) estimation problem

$$\text{maximize: } \mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})},$$

assuming a zero-mean Laplace prior on each weight θ_i, the prior is

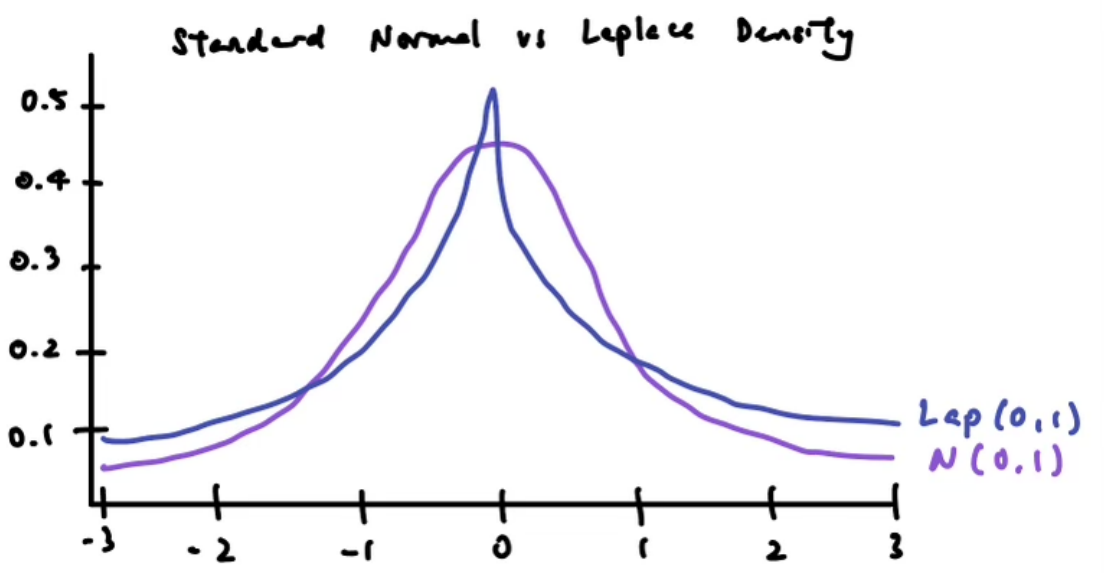
$$\mathbb{P}(\boldsymbol{\theta}) = \prod_i \frac{1}{2b}\exp\left(-\frac{|\theta_i|}{b}\right).$$

Maximizing $\mathbb{P}(\boldsymbol{\theta}|\mathcal{D})$ with this prior is equivalent to maximizing

$$\log \mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) - \lambda \sum_i |\theta_i|,$$

where $\lambda = 1/b$. This demonstrates the equivalence to ℓ_1 regularization.

Comparing the Laplace distribution $\text{Lap}(x|0,1)$ with the standard normal distribution $\mathcal{N}(x|0,1)$, the Laplace distribution has heavier tails and a sharper peak at zero. This shape makes the Laplace prior more likely to push weights exactly to zero compared to the Gaussian prior, which corresponds to ℓ_2 regularization. Thus, the Laplace prior leads to sparser solutions (graph on next page). ■


Standard Normal vs Laplace Density
0.5
0.4
0.3
0.2
0.1
-3
-2
-1
0
1
2
3
Lap(0,1)
N(0,1)