Homework 2

Due Feburary 21 at 11 pm

1. (PCA and linear regression) Consider a dataset of n 2-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^2$. Assume that the dataset is centered. Our goal is to find a line in the 2D space that lies *closest* to the data. First, we apply PCA and consider the line in the direction of the first principal direction. Second, we use least squares to fit a linear regression model where $x_i[1]$ is a feature, and $x_i[2]$ the corresponding response. Are these lines the same? Describe each line in terms of the quantity it minimizes geometrically (e.g. sum of some distance from the points to the lines), and provide an example to illustrate your description.

2. (Heartbeat) We are interested in computing the best linear estimate of the heartbeat of a fetus in the presence of strong interference in the form of the heartbeat of the baby's mother. To simplify matters, let us assume that we only want to estimate the heartbeat at a certain moment. We have available a measurement from a microphone situated near the mother's belly and another from a microphone that is away from her belly. We model the measurements as

$$\tilde{x}[1] = \tilde{b} + \tilde{m} + \tilde{z}_1 \tag{1}$$
$$\tilde{x}[2] = \tilde{m} + \tilde{z}_2, \tag{2}$$

   where $\tilde{b}$ is a random variable modeling the heartbeat of the baby, $\tilde{m}$ is a random variable modeling the heartbeat of the mother, and $\tilde{z}_1$ and $\tilde{z}_2$ model additive noise. From past data, we determine that $\tilde{b}$, $\tilde{m}$, $\tilde{z}_1$, and $\tilde{z}_2$ are all zero mean and uncorrelated with each other. The variances of $\tilde{b}$, $\tilde{z}_1$ and $\tilde{z}_2$ are equal to 1, whereas the variance of $\tilde{m}$ is much larger, it is equal to 10.

   (a) Compute the best linear estimate of $\tilde{b}$ given $\tilde{x}[1]$ in terms of MSE, and the corresponding MSE.

   (b) Compute the best linear estimate of $\tilde{b}$ given $\tilde{x}$ in terms of MSE, and the corresponding MSE.

3. (Best unbiased linear estimator) Consider the linear regression model

$$\tilde{y} = X^T\beta + \tilde{z}$$

   where $\tilde{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{p\times n}$ has rank p, $\beta \in \mathbb{R}^p$, and $\tilde{z} \in \mathbb{R}^n$ has mean zero and covariance matrix $\Sigma_z = \sigma^2 I$ for some $\sigma^2 > 0$. Here only $\tilde{z}$ and $\tilde{y}$ are random. We observe the values of $\tilde{y}$ and X and must estimate β. We consider the following question: What is the best *unbiased* linear estimator $C\tilde{y}$ of the coefficients β? Here C is any $p \times n$ deterministic matrix. An estimator is unbiased if its mean is equal to β.

   (a) What is the mean of the estimator $C\tilde{y}$?

   (b) What is the covariance matrix of $C\tilde{y}$?

   (c) Let us define $D := C - (XX^T)^{-1}X$. What must be true of D so that $C\tilde{y}$ is an unbiased estimator of β for all possible β? [Hint: Use part (a). Your answer will be a property of DX^T.]

(d) Let Σ_C denote the covariance matrix of $C\tilde{y}$ and let Σ_{OLS} denote the covariance matrix of $(XX^T)^{-1}X\tilde{y}$. Show that if $C\tilde{y}$ is an unbiased estimator of β then

$$v^T\Sigma_C v \geq v^T\Sigma_{\text{OLS}}v,$$

for all $v \in \mathbb{R}^p$. That is, least squares yields the estimator with smallest variance in any direction v. [Hint: Use part (b) to compute the covariance of $((XX^T)^{-1}X+D)\tilde{y}$ where D is defined in (c).]

(e) Now suppose that the true regression model has extra features:

$$\tilde{y} = X^T\beta + Z^T w + \tilde{z},$$

where $Z \in \mathbb{R}^{k\times n}$ and $w \in \mathbb{R}^k$. Not knowing these features, you compute the least squares estimator

$$\hat{\beta} = (XX^T)^{-1}X\tilde{y}.$$

Under what conditions on X, Z is $\hat{\beta}$ still unbiased for all possible w?

4. (Climate modeling) In this problem we model temperature trends using a linear regression model. The file `t_data.csv` contains the maximum temperature measured each month in Oxford from 1853-2014. We will use the first 150 years of data (the first $150 \cdot 12$ data points) as a training set, and the remaining 12 years as a test set.

In order to fit the evolution of the temperature over the years, we fit the following model

$$y[t] = a + bt + c\cos(2\pi t/T) + d\sin(2\pi t/T) \tag{3}$$

where $a, b, c, d \in \mathbb{R}$, $y[t]$ denotes the maximum temperature in Celsius during month t of the dataset (with t starting from 0 and ending at $162 \cdot 12 - 1$).

(a) What is the number of parameters in your model and how many data points do you have to fit the model? Are you worried about overfitting?

(b) Fit the model using least squares on the training set to find the coefficients for values of T equal to 1,2,...,20. Which of these models provides a better fit? Explain why this is the case. In the remaining question we will fix T to the value T^* that provides a better fit.

(c) Produce two plots comparing the actual maximum temperatures with the ones predicted by your model for $T := T^*$; one for the training set and one for the test set.

(d) Fit the modified model

$$y[t] = a + bt + d\sin(2\pi t/T^*) \tag{4}$$

and plot the fit to the training data as in the previous question. Explain why it is better to also include a cosine term in the model.

(e) Provide an intuitive interpretation of the coefficients a, b, c and d, and the corresponding features. According to your model, are temperatures rising in Oxford? By how much?