Mathematical Tools for Data Science Spring 2020

Sample Midterm Problems

1. *Whitening.* Consider a dataset of n centered d-dimensional vectors $x_1, x_2, \ldots, x_n$, where $n > d$. Let $u_1, \ldots, u_d$ be the principal directions of the dataset, and $\lambda_1, \ldots, \lambda_d$ the corresponding eigenvalues of the sample covariance matrix. We assume the sample covariance matrix is full rank.

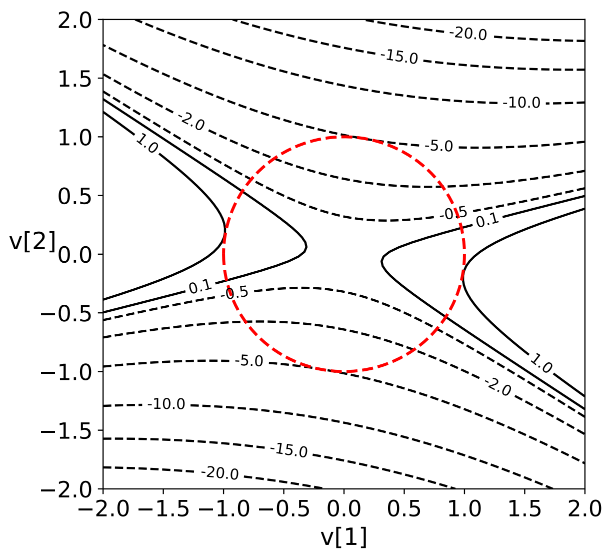
   a. If we duplicate each point, so that the data are now $x_1, x_1, x_2, x_2, \ldots, x_n, x_n$, what effect does this have on the principal directions and on the eigenvalues?

   b. Find an orthogonal matrix $A \in \mathbb{R}^{d \times d}$, such that the transformed dataset $Ax_1, Ax_2, \ldots, Ax_n$ has pairwise uncorrelated features.

   c. Find a matrix $B \in \mathbb{R}^{d \times d}$, such that the transformed dataset $Bx_1, Bx_2, \ldots, Bx_n$ has pairwise uncorrelated features and each entry $x_1[i], \ldots, x_n[i]$, $1 \le i \le d$, has unit sample variance.

   d. Would using $Bx_1, Bx_2, \ldots, Bx_n$ as features instead of the original dataset change the prediction of the response in a linear regression task?

2. *Quadratic form.* The following image shows the contour lines of the quadratic form $f(v) := v^T Av$ corresponding to a 2×2 symmetric matrix A. The unit circle is drawn in red:



   a. What are the eigenvalues of A?

   b. Can A be interpreted as a covariance matrix?

   c. Are there any points on the unit circle where the gradient of f equals zero?

3. *PCA.* We consider a dataset of d-dimensional vectors that is modeled as samples from a random vector

$$\tilde{y} := \tilde{x}v + \tilde{z}, \tag{1}$$

where $v \in \mathbb{R}^d$, $\tilde{x} \in R$ is a random variable with mean 0 and variance σ^2_{signal}, v is a fixed deterministic vector, and $\tilde{z} \in R^d$ is a Gaussian random vector with independent entries, each of which has mean zero and variance σ^2_{noise}. $\tilde{x}$ and $\tilde{z}$ are independent.

a. Sketch some samples of $\tilde{y}$ for $d = 2$ when σ_{signal} is much larger than σ_{noise}. You can assume any v for the diagram.

b. For the v you picked in part (a), sketch some samples of $\tilde{y}$ for $d = 2$ when σ_{signal} is much smaller than σ_{noise}.

c. Is averaging the dataset a good algorithm for estimating v?

d. Compute the covariance matrix of $\tilde{y}$.

e. Express the eigendecomposition of the covariance matrix in terms of σ_{signal}, σ_{noise}, v, u_2, $\ldots$, u_d. Here u_2, $\ldots$, u_d are unit ℓ_2-norm vectors that are orthogonal to v and each other.

f. Suggest an algorithm to estimate the direction of v from the data.

4. *Interference.* A radar system is trying to estimate a signal that we model as a zero-mean random variable $\tilde{y}$ with variance σ^2. Due to interference, the signal is only observed about 50% of the time. In order to improve our chances, we take two independent measurements, modeled as a 2-dimensional random vector $\tilde{x}$ with entries

$$x[i] = \begin{cases} y & \text{with probability } \frac{1}{2}, \\ \tilde{z}_i & \text{with probability } \frac{1}{2}, \end{cases} \tag{2}$$

where $\tilde{z}_1$ and $\tilde{z}_2$ are zero-mean random variables with variance σ^2 that are independent from $\tilde{y}$ and from each other. The events $\{\tilde{x}[1] = y\}$ and $\{\tilde{x}[2] = y\}$ are also independent.

a. What is the linear estimate of $\tilde{y}$ given $\tilde{x}$ that minimizes MSE?
Hint: Use the fact that for any a, b, c, and d such that $ad \neq bc$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{3}$$

Also, remember that by iterated expectation for any random variables $\tilde{a}$ and $\tilde{b}$, $\mathrm{E}(\tilde{b}) = \mathrm{E}[\mathrm{E}(\tilde{b} \,|\, \tilde{a})]$.

b. What is the corresponding MSE?

c. If $\tilde{x}[1] = \tilde{x}[2]$ we know that the estimate is perfect. Modify your estimate to return $\tilde{x}[1]$ if $\tilde{x}[1] = \tilde{x}[2]$, and otherwise return the linear estimate that minimizes MSE conditioned on $\tilde{x}[1] \neq \tilde{x}[2]$. What is the corresponding MSE?

5. *Linear regression with dimensionality reduction.* We want to fit a linear-regression model to a dataset (x_1, y_1), $\ldots$, (x_n, y_n), where $x_i \in \mathbb{R}^p$ is the ith feature vector and $y_i \in \mathbb{R}$ is the corresponding response. The number of examples is larger than the number of features, $n > p$. The features turn out to be highly correlated. The matrix of features $X \in \mathbb{R}^{p \times n}$, whose ith column equals x_i, has rank $r < p$.

a. Does the least-squares cost problem

$$\min_{\beta \in \mathbb{R}^p} \|y - X^T \beta\|_2, \tag{4}$$

where $y[i] = y_i$, have a unique solution?

b. Find a matrix $P \in \mathbb{R}^{r \times p}$ with orthonormal rows to perform dimensionality reduction on the feature vectors $x_1, x_2, \ldots, x_n$ optimally, in the sense of preserving the sample variance. Express it in terms of the SVD of $X = USV^T$, where $U \in \mathbb{R}^{p \times r}$, $S \in \mathbb{R}^{r \times r}$, and $V \in \mathbb{R}^{n \times r}$ (note that this is the reduced SVD where all singular values are nonzero).

c. Does the dimensionality reduction performed in the previous part preserve the ℓ_2 norms of the feature vectors $x_1, x_2, \ldots, x_n$ completely?

d. Assume that the data is generated by a linear model

$$y := X^T \beta_{\text{true}} + z, \tag{5}$$

where $z \in \mathbb{R}^n$ is additive noise. Explain how to fit a linear model to these data using the dimensionality-reduction matrix P so that the resulting least-squares problem has a unique solution. Write down the closed-form solution β_{LS} of the new least-squares problem in terms of the SVD of $X = USV^T$, β_{true} and z.

e. Using β_{LS} can we obtain an accurate estimate of β_{true} when z is zero? If yes, does this automatically guarantee low prediction error for new values of y? If not, does this mean that we cannot use our model to predict new values of the response?

6. *Linear regression with orthogonal features.* Consider a linear regression problem where the rows of the feature matrix X are orthogonal to each other and have unit ℓ_2 norm. The matrix of features $X \in \mathbb{R}^{p \times n}$, has it's ith column equals the i^{th} data point x_i.

a. What are the OLS coefficients equal to?

b. Express the ridge-regression estimator of the coefficients as a function of the OLS estimator and the regularization parameter λ.

c. Assume an additive model for the data,

$$\tilde{y} = X^T \tilde{\beta} + \tilde{z}, \tag{6}$$

where $\tilde{\beta}$ is a zero-mean p-dimensional random vector such that $\mathrm{E}(\|\tilde{\beta}\|_2^2) = 1$, and $\tilde{z}$ is a zero-mean Gaussian iid noise vector with variance σ^2 independent from $\tilde{\beta}$. Compute the value of λ that minimizes the mean ℓ_2-norm error $\mathrm{E}(\|\tilde{\beta}_{\text{true}} - \tilde{\beta}_{\text{RR}}\|_2^2)$, where $\tilde{\beta}_{\text{RR}}$ is the ridge-regression estimator. How does it vary with the noise variance and the number of features?