Sample Midterm Solutions

1. *Whitening.*

   a. Let Σ and Σ' be the sample covariance matrices of the original and the expanded data respectively. We have

$$\Sigma' := \frac{1}{2n}\sum_{i=1}^{n}(x_i x_i^T + x_i x_i^T) \tag{1}$$
$$= \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \tag{2}$$
$$= \Sigma. \tag{3}$$

   The covariance matrices are the same, so nothing changes.

   b. Let Σ and Σ' be the sample covariance matrices of the original and the transformed respectively. The sample covariance equals

$$\Sigma' = \frac{1}{n-1}\sum_{i=1}^{n} Ax_i(Ax_i)^T \tag{4}$$
$$= A(\frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T)A^T \tag{5}$$
$$= A\Sigma A^T \tag{6}$$
$$= AU\Lambda U^T A^T, \tag{7}$$

   where $U\Lambda U^T$ is the eigendecomposition of Σ. Setting $A = U^T$ results in a diagonal sample covariance matrix.

   c. By the argument above, setting $B := (\sqrt{\Lambda})^{-1}U^T$, where $\sqrt{\Lambda}$ is a diagonal matrix that contains the square root of the eigenvalues, results in a sample covariance matrix equal to

$$\Sigma' = BU\Lambda U^T B^T \tag{8}$$
$$= (\sqrt{\Lambda})^{-1}U^T U\Lambda U^T U(\sqrt{\Lambda})^{-1} \tag{9}$$
$$= I. \tag{10}$$

   d. It doesn't make a difference. Let $X = USV^T$ be the SVD of the original feature matrix. Then $B := S^{-1}U^T$. The OLS estimate with the modified data equals

$$y_{\text{OLS}} = (BX)^T(BXX^T B^T)^{-1}BXy \tag{11}$$
$$= X^T B^T BXy \tag{12}$$
$$= VV^T y, \tag{13}$$

   which is the exact same prediction as if we use X as the feature matrix.

2. *Quadratic form.*

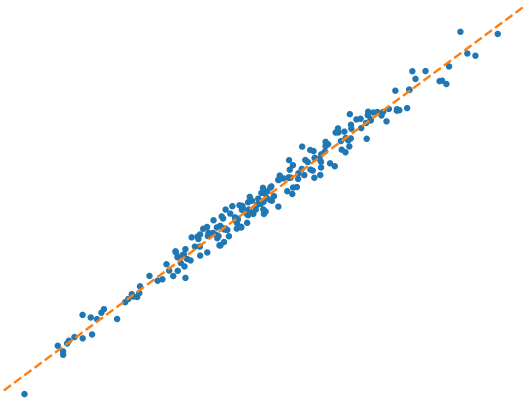
a. By the spectral theorem the eigenvalues are the maximum and minimum of the quadratic form on the unit circle. These occur at points where the contour lines are tangent to the unit circle, so $\lambda_1 = 1$ and $\lambda_2 = -5$.

b. No. The quadratic form takes negative values, which would imply that the variance in certain directions is negative, which is impossible.

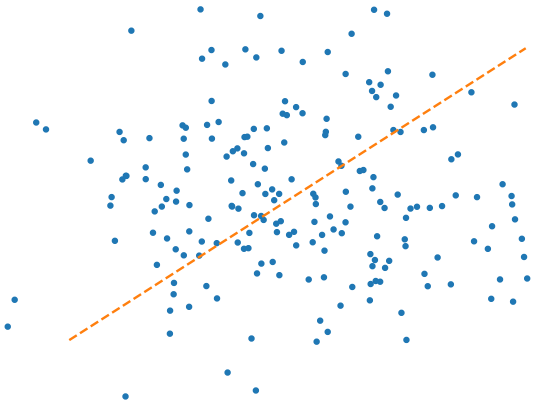
c. No. The gradient equals $2Av$, so if the gradient is zero for a nonzero v, then this implies that A has a null space. However, this is impossible because it has two nonzero eigenvalues and is consequently full rank.

3. *PCA.*

a.



b.



c. No. By linearity of expectation, the mean of the random variable that represents the data

equals

$$\begin{aligned} \mathrm{E}(\tilde{y}) &= \mathrm{E}(\tilde{x}v + \tilde{z}) && (14)\\ &= \mathrm{E}(\tilde{x})v + \mathrm{E}(\tilde{z}) && (15)\\ &= 0. && (16) \end{aligned}$$

d. By linearity of expectation and the fact that the mean of $\tilde{x}$ and $\tilde{z}$ is zero,

$$\begin{aligned} \mathrm{E}\left(\tilde{y}\tilde{y}^T\right) &= \mathrm{E}\left((\tilde{x}v + \tilde{z})(\tilde{x}v + \tilde{z})^T\right) && (17)\\ &= \mathrm{E}(\tilde{x}^2)vv^T + \mathrm{E}\left(\tilde{z}\tilde{z}^T\right) && (18)\\ &= \sigma^2_{\text{signal}}vv^T + \sigma^2_{\text{noise}}I. && (19) \end{aligned}$$

e. Setting $U := \begin{bmatrix} v/\|v\|_2 & u_2 & \cdots & u_d \end{bmatrix}$, we have $UU^T = I$ by the orthonormality assumption. We can therefore rewrite the covariance matrix as

$$\begin{aligned} \mathrm{E}\left(\tilde{y}\tilde{y}^T\right) &= \sigma^2_{\text{signal}}vv^T + \sigma^2_{\text{noise}}UU^T && (20)\\ &= U \begin{bmatrix} \sigma^2_{\text{signal}}\|v\|^2 + \sigma^2_{\text{noise}} & 0_{d-1} \\ 0_{d-1} & \sigma^2_{\text{noise}}I_{d-1} \end{bmatrix} U^T, && (21) \end{aligned}$$

where 0_{d-1} is a vector of $d-1$ zeros and I_{d-1} is the $d-1 \times d-1$ identity matrix.

f. Apply PCA and use the first principal direction as an estimate of the direction of v.

4. *Interference.*

a. Note that all random variables are zero mean. Let $\tilde{a}$ be equal to one if $\tilde{x}[i] = y$ and to zero otherwise. By iterated expectation

$$\begin{aligned} \mathrm{Var}(\tilde{x}[i]) &= \mathrm{E}(\tilde{x}[i]^2) && (22)\\ &= \mathrm{E}(\tilde{x}[i]^2 \,|\, \tilde{a} = 1)\mathrm{P}(\tilde{x}[i] = y) + \mathrm{E}(\tilde{x}[i]^2 \,|\, \tilde{a} = 0)\mathrm{P}(\tilde{x}[i] \neq y) && (23)\\ &= \frac{\sigma^2 + \sigma^2}{2} && (24)\\ &= \sigma^2. && (25) \end{aligned}$$

Similarly, let $\tilde{b}$ be equal to one if both $\tilde{x}[1]$ and $\tilde{x}[2]$ equal y and to zero otherwise. By iterated expectation

$$\begin{aligned} \mathrm{Cov}(\tilde{x}[1]\tilde{x}[2]) &= \mathrm{E}(\tilde{x}[1]\tilde{x}[2]) && (26)\\ &= \mathrm{E}(\tilde{x}[1]\tilde{x}[2] \,|\, \tilde{b} = 1)\mathrm{P}(\tilde{x}[1] = y)\mathrm{P}(\tilde{x}[2] = y) && (27)\\ &\quad + \mathrm{E}(\tilde{x}[1]\tilde{x}[2] \,|\, \tilde{b} = 0)\left(1 - \mathrm{P}(\tilde{x}[1] = y)\mathrm{P}(\tilde{x}[2] = y)\right) && (28)\\ &= \frac{\mathrm{E}(\tilde{y}^2)}{4} + 0 && (29)\\ &= \frac{\sigma^2}{4}. && (30) \end{aligned}$$

The covariance matrix equals

$$\Sigma_{\tilde{x}} = \sigma^2 \begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}. \quad (31)$$

and its inverse equals

$$\Sigma_{\tilde{x}}^{-1} = \frac{16}{15\sigma^2} \begin{bmatrix} 1 & -\frac{1}{4} \\ -\frac{1}{4} & 1 \end{bmatrix}. \tag{32}$$

The crosscovariance equals

$$\begin{aligned} \text{Cov}(\tilde{x}[i]\tilde{y}) &= \text{E}(\tilde{x}[i]\tilde{y}) && (33)\\ &= \text{E}(\tilde{x}[i]\tilde{y} \,|\, \tilde{a} = 1)\text{P}(\tilde{x}[i] = y) + \text{E}(\tilde{x}[i]\tilde{y} \,|\, \tilde{a} = 0)\text{P}(\tilde{x}[i] \neq y) && (34)\\ &= \frac{\sigma^2}{2} + 0 && (35)\\ &= \frac{\sigma^2}{2}. && (36) \end{aligned}$$

By Theorem 2.3 in the notes on linear regression the linear MMSE estimator equals

$$\begin{aligned} \hat{y}(\tilde{x}) &= \tilde{x}^T \frac{16}{15\sigma^2} \begin{bmatrix} 1 & -\frac{1}{4} \\ -\frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} \frac{\sigma^2}{2} \\ \frac{\sigma^2}{2} \end{bmatrix} && (37)\\ &= \frac{2(\tilde{x}[1] + \tilde{x}[2])}{5}. && (38) \end{aligned}$$

b. By Theorem 2.3 in the notes on linear regression the corresponding MSE equals

$$\begin{aligned} \text{MSE} &= \text{Var}(\tilde{y}) - \Sigma_{\tilde{x}\tilde{y}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{x}\tilde{y}} && (39)\\ &= \sigma^2 - \begin{bmatrix} \frac{\sigma^2}{2} \\ \frac{\sigma^2}{2} \end{bmatrix}^T \frac{16}{15\sigma^2} \begin{bmatrix} 1 & -\frac{1}{4} \\ -\frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} \frac{\sigma^2}{2} \\ \frac{\sigma^2}{2} \end{bmatrix} && (40)\\ &= \sigma^2 - \frac{2\sigma^2}{5} && (41)\\ &= \frac{3\sigma^2}{5}. && (42) \end{aligned}$$

c. Note that conditioned on $\tilde{x}[1] \neq \tilde{x}[2]$, there are only three possible events that are equally likely: $\{\tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2\}$, $\{\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y}\}$, and $\{\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2\}$. By iterated expectation

$$\begin{aligned} \text{Var}(\tilde{x}[i] \,|\, \tilde{x}[1] \neq \tilde{x}[2]) &= \text{E}(\tilde{x}[i]^2 \,|\, \tilde{x}[1] \neq \tilde{x}[2]) && (43)\\ &= \text{E}(\tilde{x}[i]^2 \,|\, \tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2)\text{P}(\tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2) && (44)\\ &\quad + \text{E}(\tilde{x}[i]^2 \,|\, \tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y})\text{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y}) && (45)\\ &\quad + \text{E}(\tilde{x}[i]^2 \,|\, \tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2)\text{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2) && (46)\\ &= \text{E}(\tilde{y}^2)\text{P}(\tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2) && (47)\\ &\quad + \text{E}(\tilde{z}_1^2)\text{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y}) && (48)\\ &\quad + \text{E}(\tilde{z}_1^2)\text{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2) && (49)\\ &= \sigma^2. && (50) \end{aligned}$$

Again, by iterated expectation

$$\begin{aligned}
\mathrm{Cov}(\tilde{x}[1]\tilde{x}[2] \,|\, \tilde{x}[1] \neq \tilde{x}[2]) &= \mathrm{E}(\tilde{x}[1]\tilde{x}[2] \,|\, \tilde{x}[1] \neq \tilde{x}[2]) && (51)\\
&= \mathrm{E}(\tilde{y}\tilde{z}_2)\mathrm{P}(\tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2) && (52)\\
&\quad + \mathrm{E}(\tilde{z}_1\tilde{y})\mathrm{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y}) && (53)\\
&\quad + \mathrm{E}(\tilde{z}_1\tilde{z}_2)\mathrm{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2) && (54)\\
&= 0. && (55)
\end{aligned}$$

The covariance matrix conditioned on $\tilde{x}[1] \neq \tilde{x}[2]$ equals

$$\Sigma_{\tilde{x} \,|\, \tilde{x}[1] \neq \tilde{x}[2]} = \sigma^2 I. \tag{56}$$

and its inverse equals

$$\Sigma^{-1}_{\tilde{x} \,|\, \tilde{x}[1] \neq \tilde{x}[2]} = \frac{1}{\sigma^2} I. \tag{57}$$

The crosscovariance equals

$$\begin{aligned}
\mathrm{Cov}(\tilde{x}[1]\tilde{y} \,|\, \tilde{x}[1] \neq \tilde{x}[2]) &= \mathrm{E}(\tilde{x}[1]\tilde{y} \,|\, \tilde{x}[1] \neq \tilde{x}[2]) && (58)\\
&= \mathrm{E}(\tilde{y}^2)\mathrm{P}(\tilde{x}[1] = \tilde{y}, \tilde{x}[2] = \tilde{z}_2) && (59)\\
&\quad + \mathrm{E}(\tilde{z}_1\tilde{y})\mathrm{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{y}) && (60)\\
&\quad + \mathrm{E}(\tilde{z}_1\tilde{y})\mathrm{P}(\tilde{x}[1] = \tilde{z}_1, \tilde{x}[2] = \tilde{z}_2) && (61)\\
&= \frac{\sigma^2}{3}. && (62)
\end{aligned}$$

By Theorem 2.3 in the notes on linear regression the linear MMSE estimator equals

$$\begin{aligned}
\hat{y}(\tilde{x}) &= \frac{1}{\sigma^2}\tilde{x}^T \begin{bmatrix} \frac{\sigma^2}{3} \\ \frac{\sigma^2}{3} \end{bmatrix} && (63)\\
&= \frac{\tilde{x}[1] + \tilde{x}[2]}{3}, && (64)
\end{aligned}$$

which is our estimate if $\tilde{x}[1] \neq \tilde{x}[2]$. The corresponding MSE conditioned on $\tilde{x}[1] \neq \tilde{x}[2]$

$$\begin{aligned}
\mathrm{E}(\tilde{y} - \,|\, \tilde{x}[1] \neq \tilde{x}[2]) &= \mathrm{Var}(\tilde{y} \,|\, \tilde{x}[1] \neq \tilde{x}[2]) - \Sigma^T_{\tilde{x}\tilde{y} \,|\, \tilde{x}[1] \neq \tilde{x}[2]} \Sigma^{-1}_{\tilde{x} \,|\, \tilde{x}[1] \neq \tilde{x}[2]} \Sigma_{\tilde{x}\tilde{y} \,|\, \tilde{x}[1] \neq \tilde{x}[2]} && (65)\\
&= \sigma^2 - \frac{1}{\sigma^2} \begin{bmatrix} \frac{\sigma^2}{3} \\ \frac{\sigma^2}{3} \end{bmatrix}^T \begin{bmatrix} \frac{\sigma^2}{3} \\ \frac{\sigma^2}{3} \end{bmatrix} && (66)\\
&= \sigma^2 - \frac{2\sigma^2}{9} && (67)\\
&= \frac{7\sigma^2}{9}. && (68)
\end{aligned}$$

Finally, the MSE equals

$$\begin{aligned}
\mathrm{E}\left((\tilde{y} - \hat{y}_{\mathrm{nl}}(\tilde{x}))^2\right) &= \mathrm{E}\left((\tilde{y} - \hat{y}_{\mathrm{nl}}(\tilde{x}))^2 \,|\, \tilde{x}[1] = \tilde{x}[2]\right) \mathrm{P}(\tilde{x}[1] = \tilde{x}[2]) && (69)\\
&\quad + \mathrm{E}\left((\tilde{y} - \hat{y}_{\mathrm{nl}}(\tilde{x}))^2 \,|\, \tilde{x}[1] \neq \tilde{x}[2]\right) \mathrm{P}(\tilde{x}[1] \neq \tilde{x}[2])\\
&= 0 + \frac{3}{4}\mathrm{E}(\tilde{y} - \hat{y}_{\mathrm{nl}}(\tilde{x}) \,|\, \tilde{x}[1] \neq \tilde{x}[2]) && (70)\\
&= \frac{7\sigma^2}{12}. && (71)
\end{aligned}$$

5. *Linear regression with dimensionality reduction.*

   a. No, let β^* be a solution. There exists at least one nonzero vector β^{null} in the null space of X^T because it is low rank. As a result, $\beta^* + \alpha\beta^{\text{null}}$ is also a solution for any value of $\alpha \in \mathbb{R}$.

   b. By Corollary 7.2 in the notes on PCA, the optimal projection is obtained by computing the eigendecomposition of the covariance matrix XX^T and using the first r principal directions. Since $XX^T = US^2U^T$, we set $P := U^T$.

   c. The ℓ_2 norm is preserved exactly. Since $X = USV^T$, we can express each data point as $x_i = Uw_i$ where $w_i := (SV^T)_i$ is the ith column of SV^T. Then we have

$$\sum_{i=1}^{n} \left|\left|U^T x_i\right|\right|_2^2 = \sum_{i=1}^{n} x_i^T U U^T x_i \tag{72}$$

$$= \sum_{i=1}^{n} w_i^T w_i \tag{73}$$

$$= \sum_{i=1}^{n} w_i^T U^T U w_i \tag{74}$$

$$= \sum_{i=1}^{n} x_i^T x_i \tag{75}$$

$$= \sum_{i=1}^{n} ||x_i||_2^2, \tag{76}$$

   because $U^T U = I$.

   d. The matrix $U^T X$ contains the points of reduced dimensionality as its columns. We fit the model by solving

$$\min_{\beta \in \mathbb{R}^r} \|y - X^T U\beta\|_2. \tag{77}$$

   $U^T X$ is full rank and has SVD SV^T, so by Eq. (54) in the lecture notes on linear regression the solution equals

$$\beta_{\text{LS}} = S^{-1}V^T y \tag{78}$$

$$= S^{-1}V^T(X^T\beta_{\text{true}} + z) \tag{79}$$

$$= U^T\beta_{\text{true}} + S^{-1}V^T z. \tag{80}$$

   e. No. We obtain $U^T\beta_{\text{true}}$ instead. Since U^T is a fat matrix, we cannot recover β_{true} from $U^T\beta_{\text{true}}$. However, we can still predict new values of y by first projecting the features x_{test} using U^T and setting

$$y_{\text{pred}} := \left\langle U^T x_{\text{test}}, \beta_{\text{LS}} \right\rangle. \tag{81}$$

6. *Linear regression with orthogonal features.*

   a. Since X has orthonormal rows, $XX^T = I$, so the OLS estimator is Xy. Each OLS coefficient is equal to the inner product of a row of X and the response vector.

b. The ridge-regression estimator equals

$$\beta_{\text{RR}} = (XX^T + \lambda I)^{-1} Xy \tag{82}$$

$$= \frac{Xy}{1+\lambda} \tag{83}$$

$$= \frac{\beta_{\text{OLS}}}{1+\lambda}. \tag{84}$$

c. The coefficients equal

$$\beta_{\text{RR}} = \frac{X\tilde{y}}{1+\lambda} \tag{85}$$

$$= \frac{XX^T\tilde{\beta} + X\tilde{z}}{1+\lambda} \tag{86}$$

$$= \frac{\tilde{\beta} + X\tilde{z}}{1+\lambda} \tag{87}$$

$$= \tilde{\beta} - \frac{\lambda\tilde{\beta}}{1+\lambda} + \frac{X\tilde{z}}{1+\lambda} \tag{88}$$

which implies

$$\text{E}(\|\tilde{\beta} - \tilde{\beta}_{\text{RR}}\|_2^2) = \text{E}\left(\left(\frac{\lambda\tilde{\beta}}{1+\lambda} - \frac{X\tilde{z}}{1+\lambda}\right)^T \left(\frac{\lambda\tilde{\beta}}{1+\lambda} - \frac{X\tilde{z}}{1+\lambda}\right)\right) \tag{89}$$

$$= \frac{\lambda^2 \text{E}(\|\tilde{\beta}\|_2^2)}{(1+\lambda)^2} + \frac{\text{E}(\tilde{z}^T X^T X\tilde{z})}{(1+\lambda)^2}. \tag{90}$$

The cross terms are zero by independence and because both random vectors have zero mean. $X\tilde{z}$ is a p-dimensional random vector with covariance matrix $X\sigma^2 I X^T = \sigma^2 I$, so $\text{E}(\tilde{z}^T X^T X\tilde{z})$ is equal to $\sigma^2 p$ by Theorem 4.5 in the notes on linear regression. We have

$$\text{E}(\|\tilde{\beta} - \tilde{\beta}_{\text{RR}}\|_2^2) = \frac{\lambda^2}{(1+\lambda)^2} + \frac{\sigma^2 p}{(1+\lambda)^2} := f(\lambda). \tag{91}$$

The derivative with respect to λ equals

$$f'(\lambda) = \frac{2\lambda}{(1+\lambda)^3} - \frac{2\sigma^2 p}{(1+\lambda)^3}, \tag{92}$$

which is zero at $\lambda = \sigma^2 p$, negative if λ is smaller and positive if λ is larger. The optimal λ therefore equals $\sigma^2 p$. It is proportional to the noise variance and to the number of features.