Midterm Solutions

1. *Short questions.*

   a. No. Projecting onto the direction of maximum variation will make the two clusters indistinguishable.

   b. The covariance matrix of $A\tilde{x}$ equals

   $$\Sigma_{A\tilde{x}} = A\Sigma_{\tilde{x}}A^T \tag{1}$$
   $$= AA^T. \tag{2}$$

   To find the singular values we compute the eigendecomposition of this covariance matrix and then take a square root.

   c. If some of the eigenvalues of the eigendecomposition of the feature matrix are very small, then OLS will suffer from noise amplification and it is probably a good idea to apply ridge regression.

   d. No. Let X be the feature matrix and y the response. Let β_{OLS} and β_{RR} be the OLS and ridge-regression coefficient estimates respectively. By definition, the training error of OLS equals

   $$||y - X\beta_{\text{OLS}}||_2^2 = \min_{\beta} ||y - X\beta||_2^2 \tag{3}$$
   $$\leq ||y - X\beta_{\text{RR}}||_2^2, \tag{4}$$

   for any value of λ.

2. *Normalization.*

   a. We have $\tilde{y} = A\tilde{x}$, where

   $$A := \begin{bmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{20} & 0 \\ 0 & 0 & \frac{1}{0.4} \end{bmatrix}, \tag{5}$$

   so

   $$\Sigma_{\tilde{y}} := A\Sigma_{\tilde{x}}A^T \tag{6}$$
   $$= \begin{bmatrix} 1 & 0.125 & 0 \\ 0.125 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{7}$$

   b. No. For example, in the direction of the vector

   $$v := \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix} \tag{8}$$

   the directional variance of $\tilde{y}$ equals

   $$\text{Var}\left(v^T\tilde{y}\right) = v^T\Sigma_{\tilde{y}}v \tag{9}$$
   $$= 1.125. \tag{10}$$

c. The eigendecomposition of $\Sigma_{\tilde{x}}$ yields eigenvalues equal to 402, 97.9 and 0.16, corresponding to the eigenvectors

$$u_1(\tilde{x}) = \begin{bmatrix} -0.997 \\ 0.082 \\ 0 \end{bmatrix}, \quad u_2(\tilde{x}) = \begin{bmatrix} -0.082 \\ -0.997 \\ 0 \end{bmatrix}, \quad u_3(\tilde{x}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{11}$$

respectively. To reduce dimensionality we compute the inner product with $u_1(\tilde{x})$ and $u_2(\tilde{x})$.

The eigendecomposition of $\Sigma_{\tilde{y}}$ yields eigenvalues equal to 1.125, 1 and 0.875, corresponding to the eigenvectors

$$u_1(\tilde{y}) = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix}, \quad u_2(\tilde{y}) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad u_3(\tilde{y}) = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{bmatrix} \tag{12}$$

respectively. To reduce dimensionality we compute the inner product with $u_1(\tilde{y})$ and $u_2(\tilde{y})$.

d. (1) Using $\tilde{y}$ is probably better because the variances and covariances correspond to different units that are not directly comparable. Normalizing makes the analysis independent of the units. The third feature is height in meters so it is normal that its variance is smaller than one. This does not mean that it is negligible.

(2) Using $\tilde{x}$ is probably better. The three features are dimensions of cars in the same unit. The variance of the height is very small compared to the other two, which indicates that the height of the cars does not vary at all. It therefore makes sense to ignore it while reducing the dimensionality.

3. *Noise cancellation.*

a. By independence

$$\begin{aligned} \text{Var}(\tilde{x}[1]) &= \text{Var}(\tilde{y}) + \alpha^2 \text{Var}(\tilde{z}) && (13) \\ &= 1 + 100\alpha^2, && (14) \\ \text{Cov}(\tilde{y}\tilde{x}[1]) &= 1 && (15) \end{aligned}$$

so

$$\begin{aligned} \Sigma_{\tilde{x}[1]} &= 1 + 100\alpha^2, && (16) \\ \Sigma_{\tilde{y}\tilde{x}[1]} &= 1 && (17) \end{aligned}$$

and by Theorem 2.3 in the notes the estimate equals

$$\hat{y}(\tilde{x}[1]) = \frac{\tilde{x}[1]}{1 + 100\alpha^2}, \tag{18}$$

and the corresponding MSE equals

$$\begin{aligned} \text{E}((\hat{y}(\tilde{x}[1]) - \tilde{b})^2) &= \text{Var}(\tilde{y}) - \frac{1}{1 + 100\alpha^2} && (19) \\ &= \frac{100\alpha^2}{1 + 100\alpha^2}. && (20) \end{aligned}$$

The estimate just shrinks the signal.

b. There are two ways to find the answer. We can notice that we can express $\tilde{y}$ exactly as a linear combination of the measurements:

$$\hat{y}(\tilde{x}) := \frac{\tilde{x}[1] - \alpha\tilde{x}[2]}{1-\alpha^2} \tag{21}$$

$$= \frac{\tilde{y} + \alpha\tilde{z} - \alpha(\alpha\tilde{y} + \tilde{z})}{1-\alpha^2} \tag{22}$$

$$= \tilde{y}. \tag{23}$$

This means that the linear estimate is perfect so the corresponding MSE equals zero. The linear estimate cancels out the noise by scaling the second measurement and subtracting it from the first one.

The second way to answer is by using the formula for the best linear estimator in Theorem 2.3 of the notes on linear regression. By independence

$$\mathrm{Var}(\tilde{x}[1]) = 1 + 100\alpha^2, \tag{24}$$

$$\mathrm{Var}(\tilde{x}[2]) = \alpha^2\mathrm{Var}(\tilde{y}) + \mathrm{Var}(\tilde{z}) \tag{25}$$

$$= \alpha^2 + 100, \tag{26}$$

$$\mathrm{Cov}(\tilde{x}[1]\tilde{x}[2]) = \alpha\mathrm{E}(\tilde{y}^2) + \alpha\mathrm{E}(\tilde{z}^2) \tag{27}$$

$$= 101\alpha, \tag{28}$$

$$\mathrm{Cov}(\tilde{y}\tilde{x}[1]) = 1 \tag{29}$$

$$\mathrm{Cov}(\tilde{y}\tilde{x}[2]) = \alpha \tag{30}$$

$$\tag{31}$$

so

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1+100\alpha^2 & 101\alpha \\ 101\alpha & \alpha^2+100 \end{bmatrix}, \tag{32}$$

$$\Sigma_{\tilde{b}\tilde{x}} = \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \tag{33}$$

and by Theorem 2.3 in the notes the estimate equals

$$\hat{y}(\tilde{x}) = \tilde{x}^T \begin{bmatrix} 1+100\alpha^2 & 101\alpha \\ 101\alpha & \alpha^2+100 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \tag{34}$$

$$= \tilde{x}^T \frac{1}{(1+100\alpha^2)(\alpha^2+100) - 101^2\alpha^2} \begin{bmatrix} \alpha^2+100 & -101\alpha \\ -101\alpha & 1+100\alpha^2 \end{bmatrix} \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \tag{35}$$

$$= \tilde{x}^T \frac{1}{100(1-\alpha^2)^2} \begin{bmatrix} 100(1-\alpha^2) \\ -100\alpha(1-\alpha^2) \end{bmatrix} \tag{36}$$

$$= \frac{\tilde{x}[1] - \alpha\tilde{x}[2]}{1-\alpha^2}. \tag{37}$$

4. *Adversarial noise.*

a. Let $X = USV^T$ be the SVD of the feature matrix. For any z,

$$\beta_{\mathrm{OLS}} = US^{-1}V^T y, \tag{38}$$

$$\beta^{\mathrm{mod}}_{\mathrm{OLS}}(z) = US^{-1}V^T(y+z), \tag{39}$$

so

$$z_{\text{OLS}}^{\text{adv}} := \arg \max_{||z||_2 \leq \gamma} \left|\left|\beta_{\text{OLS}} - \beta_{\text{OLS}}^{\text{mod}}(z)\right|\right|_2^2 \tag{40}$$

$$= \arg \max_{||z||_2 \leq \gamma} \left|\left|US^{-1}V^T z\right|\right|_2^2 . \tag{41}$$

Let s_p correspond to the smallest singular value of X and v_p the corresponding right singular vector. The largest singular value of $US^{-1}V^T$ is $1/s_p$ and the corresponding right singular vector is v_p. By the properties of the SVD we have

$$v_p = \arg \max_{||v||_2 \leq 1} \left|\left|US^{-1}V^T v\right|\right|_2^2 , \tag{42}$$

so the z that produces the maximum perturbation is $z_{\text{OLS}}^{\text{adv}} = \gamma v_p$. Note that if the smallest singular value is not unique, we can choose any linear combination of the right singular vectors corresponding to the smallest singular values.

b. If $\frac{1}{n}XX^T$ is a good approximation to the covariance matrix of $\tilde{x}_{\text{test}}$ then v_p would be aligned with the direction of least variance of the test vector, so the variance of the perturbation will probably be larger if it is *not* a good approximation.

c. For any z, by Theorem 5.2 in the notes on linear regression

$$\beta_{\text{RR}}(\lambda) = (XX^T + \lambda I)^{-1} Xy \tag{43}$$

$$= U(S^2 + \lambda I)^{-1} SV^T y, \tag{44}$$

$$\beta_{\text{RR}}^{\text{mod}}(z, \lambda) = U(S^2 + \lambda I)^{-1} SV^T (y + z), \tag{45}$$

so

$$z_{\text{RR}}^{\text{adv}}(\lambda) := \arg \max_{||z||_2 \leq \gamma} \left|\left|\beta_{\text{RR}}(\lambda) - \beta_{\text{RR}}^{\text{mod}}(z)\right|\right|_2^2 \tag{46}$$

$$= \arg \max_{||z||_2 \leq \gamma} \left|\left|U(S^2 + \lambda I)^{-1} SV^T z\right|\right|_2^2 . \tag{47}$$

By the same argument as the previous question, the maximum perturbation is achieved by the right singular vector v_{j^*} corresponding to the index j^* defined as

$$j^* := \arg \max_j \frac{s_j}{s_j^2 + \lambda}. \tag{48}$$

The squared ℓ_2 norm of the perturbation is equal to

$$\gamma^2 \left(\frac{s_j}{s_j^2 + \lambda}\right)^2 \leq \frac{\gamma^2}{s_j^2} \tag{49}$$

$$\leq \frac{\gamma^2}{s_p^2}, \tag{50}$$

which is the perturbation achieved for OLS. The perturbation for RR is therefore smaller.

d. If we set λ extremely large, the ridge-regression coefficients will be close to zero, for any values of y and z. This is not a reasonable option, because the learned coefficients are useless and not related to the data.