DS-GA.1013 Mathematical Tools for Data Science
Homework 4
Yves Greatti - yg390

1. (Condition number) Let $A \in \mathbb{R}^{n\times n}$ be invertible, and let $x_{\text{true}}, y \in \mathbb{R}^n$ satisfy $Ax_{\text{true}} = y$. We are interested in what happens if $y$ is perturbed additively by a vector $z \in \mathbb{R}^n$, i.e. if we solve

$$Aw = y + z. \tag{1}$$

(a) The operator norm of a matrix $M$ is equal to

$$||M|| := \max_{||v||_2=1} ||Mv||_2 \,, \tag{2}$$

which we know is equal to the maximum singular value. What is the operator norm of $A^{-1}$?

First A is invertible if and only if $Ax = 0$ admits the only solution $x = 0$ (if not $Ker(A) \neq 0$ and A is not full row rank). $A$ being symmetric by the spectral theorem, we have $A = USU^T$ and we can compute an SVD of A by $A = U|S|U^T$. Then 0 cannot be an eigenvalue or a singular value, and A has only positive singular values, let $s_1 \geq \ldots \geq s_n > 0$, these singular values. A is invertible so $AA^{-1} = I$, let the SVD of A be $USV^T$ then $AA^{-1} = USV^TA^{-1} = I$. $U, V$ being orthonormal matrices, and multiplying on both sides by $U^T, S^{-1}, V$, we have $A^{-1} = VS^{-1}U^T$. Therefore the singular values of $A^{-1}$ are $\frac{1}{s_n} \geq \ldots \geq \frac{1}{s_1}$ thus $||A^{-1}|| = \max_{||v||_2=1} ||A^{-1}v||_2 = \frac{1}{s_n}$.

(b) Prove that $\|w - x_{\text{true}}\|_2 \leq \|z\|_2/s_n$, where $s_j$ denotes the $j$th singular value of $A$.

*Proof.* $A(w - x_{\text{true}}) = Aw - Ax_{\text{true}} = y + z - y = z$. Hence

$$\|A^{-1}A(w - x_{\text{true}})\|_2 = \|w - x_{\text{true}}\|_2 = \|A^{-1}z\|_2$$

$L_2$ norm being consistent $\|A^{-1}z\|_2\| \leq ||A^{-1}||\,\|z\|_2 \leq \|z\|_2/s_n$.

□

(c) If $x_{\text{true}} \neq 0$ prove that

$$\frac{\|w - x_{\text{true}}\|_2}{\|x_{\text{true}}\|_2} \leq \kappa(A)\frac{\|z\|_2}{\|y\|_2}.$$

Here $\kappa(A) := s_1/s_n$ is called the *condition number* of $A$.

Similarly to part (b), we have

$$
\begin{aligned}
\|A^{-1}Ax_{\text{true}}\|_2 &= \|A^{-1}y\|_2 \\
\|x_{\text{true}}\|_2 &= \|A^{-1}y\|_2 \\
\|x_{\text{true}}\|_2 &\leq \left|\left|A^{-1}\right|\right| \|y\|_2 \\
&= \frac{\|y\|_2}{s_n}
\end{aligned}
$$

So for $x_{\text{true}} \neq 0$, we now have

$$
\begin{aligned}
\frac{\|w - x_{\text{true}}\|_2}{\|x_{\text{true}}\|_2} &\leq \frac{\|z\|_2}{s_n} \frac{s_n}{\|y\|_2} \\
&\leq \frac{\|z\|_2}{s_n} \frac{s_1}{\|y\|_2} \quad \text{since } s_n \leq s_1 \\
&= \kappa(A) \frac{\|z\|_2}{\|y\|_2}
\end{aligned}
$$

2. (Simple linear regression) We consider a linear model with one feature ($p := 1$). The data are given by

$$\tilde{y}_i := x_i\beta + \tilde{z}_i, \quad 1 \leq i \leq n, \tag{3}$$

where $\beta \in \mathbb{R}$, $x_i \in \mathbb{R}$, and $\tilde{z}_1, \ldots, \tilde{z}_n$ are iid Gaussian random variables with zero mean and variance $\sigma^2$. A reasonable definition of the *energy* in the feature is its sample mean square $\gamma^2 := \frac{1}{n}\sum_{i=1}^n x_i^2$. We define the signal-to-noise ratio in the data as SNR:$= \gamma^2/\sigma^2$.

(a) What is the distribution of the OLS estimate $\tilde{\beta}_{OLS}$ as a function of the SNR?

The OLS estimate $\tilde{\beta}_{OLS}$ is given by $\tilde{\beta}_{OLS} = \arg\min_\beta \sum_{i=1}^n (\tilde{y}_i - \beta x_i)^2 = f(\beta)$. The function $f$ is a quadratic form. Its first derivative and second derivative are:

$$f'(\beta) = \frac{d}{d\beta}\sum_{i=1}^n (\tilde{y}_i - \beta x_i)^2 = -2\sum_{i=1}^n x_i(\tilde{y}_i - \beta x_i) = -2\sum_{i=1}^n x_i\tilde{y}_i + 2\beta\sum_{i=1}^n x_i^2$$

$$f''(\beta) = 2\sum_{i=1}^n x_i^2$$

if $f'(\beta) = 0$ means $\beta = \tilde{\beta}_{OLS}$ and $f'(\beta) = 0 \Rightarrow \tilde{\beta}_{OLS} = \frac{\sum_{i=1}^n x_i\tilde{y}_i}{\sum_{i=1}^n x_i^2}$.

$$\begin{aligned}
\tilde{\beta}_{OLS} &= \frac{\sum_{i=1}^n x_i\tilde{y}_i}{\sum_{i=1}^n x_i^2} \\
&= \frac{\sum_{i=1}^n x_i(x_i\beta + \tilde{z}_i)}{\sum_{i=1}^n x_i^2} \\
&= \beta\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} + \frac{\sum_{i=1}^n x_i\tilde{z}_i}{\sum_{i=1}^n x_i^2} \\
&= \beta + \frac{\sum_{i=1}^n x_i\tilde{z}_i}{\sum_{i=1}^n x_i^2}
\end{aligned}$$

$\tilde{\beta}_{OLS}$ is a Gaussian random variable (being a linear of a Gaussian variable) with mean, using linearity of the expectation:

$$\mathrm{E}[\tilde{\beta}_{OLS}] = \mathrm{E}[\beta + \frac{\sum_{i=1}^n x_i\tilde{z}_i}{\sum_{i=1}^n x_i^2}] = \mathrm{E}[\beta] + \mathrm{E}[\frac{\sum_{i=1}^n x_i\tilde{z}_i}{\sum_{i=1}^n x_i^2}] = \beta + \frac{\sum_{i=1}^n x_i\mathrm{E}[\tilde{z}_i]}{\sum_{i=1}^n x_i^2} = \beta$$

where for the last equality we have used that $\tilde{z}_n = i, i = 1, \ldots, n$ are Gaussian random

variables with zero mean. And the variance is:

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}_{OLS}) &= \mathrm{Var}(\beta + \frac{\sum_{i=1}^n x_i \tilde{z}_i}{\sum_{i=1}^n x_i^2}) \\
&= \mathrm{Var}(\frac{\sum_{i=1}^n x_i \tilde{z}_i}{\sum_{i=1}^n x_i^2}) \text{ since variance is invariant to change to the location parameter} \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \mathrm{Var}(\sum_{i=1}^n x_i \tilde{z}_i) \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n \mathrm{Var}(x_i \tilde{z}_i) \quad z_i \text{ being iid} \\
&= \frac{1}{(\sum_{i=1}^n x_i^2)^2} \sum_{i=1}^n x_i^2 \mathrm{Var}(\tilde{z}_i) \\
&= \frac{\sigma^2}{(\sum_{i=1}^n x_i^2)} = \frac{\sigma^2}{n\gamma^2} = \frac{1}{n\,\mathrm{SNR}}
\end{aligned}$$

(b) If the SNR is fixed, how does the estimate behave as $n \to \infty$? If $n$ is fixed, how does the estimate behave as $\mathrm{SNR} \to \infty$? Can this behavior change if the noise is iid, has zero mean and variance $\sigma^2$, but is not Gaussian? Prove that it doesn't or provide an example where it does.

If the SNR is fixed then $\mathrm{Var}(\tilde{\beta}_{OLS}) \to_{n\to\infty} 0$ and similarly for $n$ fixed, $\mathrm{Var}(\tilde{\beta}_{OLS}) \to_{\mathrm{SNR}\to\infty} 0$, which means that we have an infinite number of samples or there is considerably more signal into the data than noise, the estimate becomes unbiased and is equal to true $\beta$. In our proof above there was no reference to the specifics of the Gaussian like the pdf or cdf only that we have a deterministic mean and variance. The behavior does not change if the noise is iid, has zero mean and variance $\sigma^2$ but is not Gaussian.

(c) Can the behavior of the estimator as $n \to \infty$ change if the noise is not iid? Prove that it doesn't or provide a counterexample.

In our proof above each of the noise $z_i, i = 1, \ldots, n$ variable being independent was essential when we expanded the sum of the variances related to the noise. As a counter-example: "Imagine that you wanted to learn about abilities of children in a classroom, so you give them some tests. You could use the test results as an indicator of the abilities of kids only if they did them by themselves, independently of each other. If they interacted then you'd probably measure abilities of the most clever kid, or the most influential one. The kids also need to be "identically distributed", so they cannot come from different countries, speak different languages, be in different ages since it will make it hard to interpret the results (maybe they did not understand the questions and answered randomly). You can deal with non-i.i.d. data but then you have to worry about "noise" in your data much more."

3. (Best unbiased 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 $\beta$. Consider a linear estimator of the form $C\tilde{y}$ where $C \in \mathbb{R}^{p\times n}$ (note that $X$ and $C$ are both deterministic, i.e., not random).

(a) What is the mean $\mu = \mathrm{E}[C\tilde{y}]$?

*Proof.* By linearity of the expectation and using the assumption that $\tilde{z} \in \mathbb{R}^n$ has mean zero, we have:

$$\begin{aligned}\mathrm{E}[C\tilde{y}] &= \mathrm{E}[CX^T\beta + C\tilde{z}]\\ &= CX^T\beta + C\mathrm{E}[\tilde{z}]\\ &= CX^T\beta = \mu\end{aligned}$$

□

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

$$\mathrm{E}[(C\tilde{y})(C\tilde{y})^T] - \mu\mu^T.$$

*Proof.* First by linearity of the expectation applied to each row of the matrix $C$ we write:

$$\mathrm{E}([C\tilde{y})(C\tilde{y})^T] = C\mathrm{E}[\tilde{y}\tilde{y}^T]C^T$$

And expanding the covariance of $\tilde{y}$, we note that:

$$\begin{aligned}\mathrm{E}[\tilde{y}^T\tilde{y}] &= \mathrm{E}[(X^T\beta + \tilde{z})^T(X^T\beta + \tilde{z})]\\ &= \mathrm{E}[(\beta^T X + \tilde{z}^T)(X^T\beta + \tilde{z})]\\ &= \mathrm{E}[\beta^T XX^T\beta I + 2\beta^T X\tilde{z} + \tilde{z}^T\tilde{z}]\\ &= \mathrm{E}[\beta^T XX^T\beta I] + 2\beta^T X\mathrm{E}[\tilde{z}] + E[\tilde{z}^T\tilde{z}]\\ &= \beta^T XX^T\beta I + \sigma^2 I\\ &= (\|\beta^T X\|_2^2 + \sigma^2)I\end{aligned}$$

where we used on the third line linearity of the expectation and the assumption that $\mathrm{E}[\tilde{z}] = 0$. Thus $\mathrm{E}([C\tilde{y})(C\tilde{y})^T] = (\|\beta^T X\|_2^2 + \sigma^2)CC^T$ and $\mathrm{Cov}(C\tilde{y}, C\tilde{y}) = (\|\beta^T X\|_2^2 + \sigma^2)CC^T - \mu\mu^T$. □

(c) Write $C = (XX^T)^{-1}X + D$ for some $D \in \mathbb{R}^{p\times n}$. What must be true of $D$ so that $C\tilde{y}$ is an unbiased estimator of $\beta$ for all possible $\beta$? That is, what must be true so that $\mathrm{E}[C\tilde{y}] = \beta$ for all $\beta$? [Hint: Use part (a). Your answer will be a property of $DX^T$.]

$C\tilde{y}$ is an unbiased estimator of $\beta$ if $\mathrm{E}[C\tilde{y}] = \beta, \forall\beta$. Using part (a) $\mathrm{E}[C\tilde{y}] = CX^T\beta$ and this has to be verified for any $\beta$, let $\beta = \mathbf{1} \in \mathbb{R}^p$ then we want $CX^T\mathbf{1}_{\mathbb{R}^p} = \mathbf{1}_{\mathbb{R}^p}$. Substituting the value of $C$, we obtain:

$$\begin{aligned} CX^T\mathbf{1}_{\mathbb{R}^p} &= [(XX^T)^{-1}X + D]X^T\mathbf{1}_{\mathbb{R}^p} \\ &= (XX^T)^{-1}XX^T\mathbf{1}_{\mathbb{R}^p} + DX^T\mathbf{1}_{\mathbb{R}^p} \\ &= \mathbf{1}_{\mathbb{R}^p} + DX^T\mathbf{1}_{\mathbb{R}^p} \end{aligned}$$

For $C\tilde{y}$ to be an unbiased estimator of $\beta$, we need $\mathbf{1}_{\mathbb{R}^p} + DX^T\mathbf{1}_{\mathbb{R}^p} = \mathbf{1}_{\mathbb{R}^p} \Rightarrow DX^T\mathbf{1}_{\mathbb{R}^p} = 0_{\mathbb{R}^p} \Rightarrow DX^T = 0_{p\times p}$ or $XD^T = 0$. The span of column space of $D^T$ has to be orthogonal to the row space of $X$ (which is the feature space).

(d) Let $\Sigma_C$ denote the covariance matrix of $C\tilde{y}$ and let $\Sigma_{\text{OLS}}$ denote the covariance matrix of $(XX^T)^{-1}X\tilde{y}$. Show that if $C\tilde{y}$ is an unbiased estimator of $\beta$ 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 $((XX^T)^{-1}X + D)\tilde{y}$.]

Using part (b) we first compute $CC^T$:

$$\begin{aligned} CC^T &= [(XX^T)^{-1}X + D][(XX^T)^{-1}X + D]^T \\ &= (XX^T)^{-1}XX^T(XX^T)^{-1} + 2DX^T(XX^T)^{-1} + DD^T \\ &= (XX^T)^{-1}XX^T(XX^T)^{-1} + DD^T \end{aligned}$$

where we have used the assumption that $C\tilde{y}$ is an unbiased estimator of $\beta$ for the last equality. Using part (b) we know that $\Sigma_{\text{OLS}} = (\|\beta^T X\|_2^2 + \sigma^2)(XX^T)^{-1}XX^T(XX^T)^{-1} - \mu\mu^T$. Putting all together:

$$\begin{aligned} v^T\Sigma_C v &= v^T(XX^T)^{-1}XX^T(XX^T)^{-1}v + v^T DD^T v - v^T\mu\mu^T v \\ &= v^T(XX^T)^{-1}XX^T(XX^T)^{-1}v - v^T\mu\mu^T v + \|D^T v\|_2^2 \\ &= v^T\Sigma_{\text{OLS}}v + \|D^T v\|_2^2 \\ &\geq v^T\Sigma_{\text{OLS}}v \quad \text{since } \|D^T v\|_2^2 \geq 0 \end{aligned}$$

(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 $w$?

$\hat{\beta}, \forall w$ is unbiased estimator of $\beta$ if $\mathrm{E}[\hat{\beta}] = \beta, \forall \beta$, which means that $\mathrm{E}[(XX^T)^{-1}X\tilde{y}] = \beta$, however be substitution $\tilde{y}$, we work out that:

$$\begin{aligned}
\mathrm{E}[(XX^T)^{-1}X(X^T\beta + \tilde{z}) + (XX^T)^{-1}XZ^Tw] &= \mathrm{E}[(XX^T)^{-1}X(X^T\beta + \tilde{z})] + \mathrm{E}[(XX^T)^{-1}XZ^Tw] \\
&= \beta + (XX^T)^{-1}X\mathrm{E}[\tilde{z}] + (XX^T)^{-1}XZ^Tw \\
&= \beta + (XX^T)^{-1}X\ 0 + (XX^T)^{-1}XZ^Tw \\
&= \beta + (XX^T)^{-1}XZ^Tw
\end{aligned}$$

which implies that $(XX^T)^{-1}XZ^Tw = 0_{\mathbb{R}^p}, \forall w \in \mathbb{R}^k$. Since this has to be verified for any $w$, by selecting a vector $w_i$ where $i, i = 1, \ldots, k$ is the only non zero coordinate equal to 1, we find that we have an unbiased estimator $\hat{\beta}$ for $\tilde{y}$ if the span of column space of $Z^T$ is orthogonal to the row space of $X$ (which is the feature space).

4. (Distribution of $\beta$) In this question, we will investigate how the coefficients of regression, $\beta$ is distributed. We will use the combined cycle power plant data set to regress for the net hourly electrical energy output as a function of the ambient temperature and exhaust vacuum. The support code loads the datasets and defines these subset of variables as $X$ and $y$ respectively. We will fit a regression to obtain $\beta_0, \beta$ which minimizes $y = \beta_0 + \beta^T x$.

   To study the distribution of $\beta$, we split our dataset into $500$ bootstrap samples, each with $100$ data points. We fit linear regression individually on each of these $500$ bootstrap samples to obtain $\beta^1, \beta^2, \ldots, \beta^{500}$.

   (a) Plot a histogram of the distribution of $\beta_1^k$ and $\beta_2^k$ where $k$ refers to the $k^{th}$ bootstrap sample and $\beta_i$ refers to the $i^{th}$ component of $\beta$. The support code handles the actual plotting part, you only have to compute the $\beta^k$s.

   (b) Make a scatter plot of $\beta_1^k$ vs $\beta_2^k$. Plot the principal directions of the actual data $X$ and the principal directions of $\beta^k$s.

   (c) Do the principal directions of $X$ datapoints and $\beta^k$ datapoints align? Give a condition on the data generation process under which these principal directions will align.