1 Frobenius Norm, Trace, SVD

The Frobenius norm of a matrix is defined as the square root of the sum of the absolute squares of its elements:

$$\|M\|_F = \sqrt{\sum_{i,j} |m_{i,j}|^2}$$

In class we have used the following property to simplify our calculations of the Frobenius norm,

$$\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)}$$

(a) Prove that the above equation is true.

Solution:

$$\|M\|_F^2 = \sum_{i,j} |m_{i,j}|^2 = \sum_{i=1}^{n} (MM^T)_{i,i}$$
$$\|M\|_F^2 = \mathrm{Tr}(MM^T)$$
$$\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)}$$

(b) Now show that the Frobenius norm of a square matrix is equal to the square root of the sum of the singular values:

$$\|M\|_F = \sqrt{\sum_i \sigma_i^2}$$

Solution:

$$\begin{aligned}
\|M\|_F^2 = \mathrm{Tr}(M^T M) &= \mathrm{Tr}((V\Sigma^T U^T)(U\Sigma V^T)) \\
&= \mathrm{Tr}(V\Sigma^T \Sigma V^T) && UU^T = I \\
&= \mathrm{Tr}(V\Sigma^2 V^T) && \Sigma \text{ is a diagonal matrix} \\
&= \mathrm{Tr}(\Sigma^2 V^T V) && \text{by the cyclic property of trace} \\
&= \mathrm{Tr}(\Sigma^2) && V^T V = I \\
&= \sum_i \sigma_i^2
\end{aligned}$$

(c) In HW 4 we used the Eckart-Young-Mirsky thereom to find the closest lower-rank approximation of a degenerate matrix in the Frobenius Norm. Show that the following is true:

Given a matrix $M \in R^{mxn}$ with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$

$$\min_{\tilde{M}_k} \|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^{m} \sigma_i^2$$

$Hint$: First show that i) There is a rank k matrix $\tilde{M}_k$ such that

$$\|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

then show that ii) $\tilde{M}_k$ yields the minimal solution to the optimization problem.

Solution:

Claim: i) There is a rank k matrix $\tilde{M}_k$ such that

$$\|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

Proof: i)

$$\tilde{M}_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

$$\|M - \tilde{M}_k\|_F^2 = \|\sum_{i=1}^{m} \sigma_i u_i v_i^T - \sum_{i=1}^{k} \sigma_i u_i v_i^T\|_F^2 = \|\sum_{i=k+1}^{m} \sigma_i u_i v_i^T\|_F^2$$

$$= \sum_{i=k+1}^{m} \sigma_i^2$$

Claim: ii) For any rank k matrix $\tilde{M}_k$

$$\|M - \tilde{M}_k\|_F^2 \geq \sum_{i=k+1}^{n} \sigma_i^2$$

Proof: ii) What we really want to show is the following:
For any matrix B of rank at most k

$$\|A - A_k\|_F \leq \|A - B\|_F$$

Let B minimize $\|A - B\|_F^2$ among all rank k or less matrices. Let V be the space spanned by the rows of B. V is at most rank k, since B is a rank k matrix. If B minimizes $\|A - B\|_F^2$, then it must be that each row of B is the projection of the corresponding row of A onto V. If this were not true, we could simply replace a row of B with the projection of the corresponding row of A onto V, lowering $\|A - B\|_F^2$ while not changing V. Since each row of B is the projection of the corresponding row of A onto V, $\|A - B\|_F^2$ is the sum of squared distances of rows of A to V. Since A_k minimized the sum of squared distance of rows fo A to any k-d subspace, it follows that $\|A - A_k\|_F \leq \|A - B\|_F$

2 Bias/Variance for K-Nearest Neighbors Regression

Suppose we have n training points x_i with labels y_i. We want to model a regression problem with k-nearest neighbors regression. K-nearest neighbors works as follows: for a particular data point z, the k-nearest neighbors regression algorithm finds the closest k points to z in our n training points and predicts the value label for z by averaging the labels of the closest k points. More formally, we model our hypothesis $h(z)$ as

$$h(z) = \frac{1}{k}\sum_{i=1}^{n} N(x_i, z, k)$$

where the function N is defined as

$$N(x_i, z, k) = \begin{cases} y_i & \text{if } x_i \text{ is one of the } k \text{ closest points to } z \\ 0 & o.w. \end{cases}$$

Suppose also we assume our labels $y_i = f(x_i) + \varepsilon$, where ε is the noise that comes from $\mathcal{N}(0, \sigma^2)$ and f is the true function.

(a) Derive the bias^2 of our model for given x_i, y_i pairs. Remember that the bias is simply $(\mathbb{E}(h(z)) - f(z))^2$.

Solution: Let $x_1 \ldots x_k$ be the k closest points.

$$\begin{aligned}(\mathbb{E}(h(z)) - f(z))^2 &= (\mathbb{E}(\frac{1}{k}\sum_{i=1}^{n} N(x_i, z, k)) - f(z))^2 = (\mathbb{E}(\frac{1}{k}\sum_{i=1}^{k} y_i) - f(z))^2 \\ &= (\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}(y_i) - f(z))^2 = (\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}(f(x_i) + \varepsilon) - f(z))^2 \\ &= (\frac{1}{k}\sum_{i=1}^{k} f(x_i) - f(z))^2\end{aligned}$$

(b) How well does k-nearest neighbors behave as $k \longrightarrow \infty$? How about when $k = 1$? Comment.

Solution: When $k \longrightarrow \infty$, then $\frac{1}{k}\sum_{i=1}^{k} f(x_i)$ goes to the average label for x. When $k = 1$, then the bias is simply $f(x_1) - f(z)$. Assuming x_1 is close enough to $f(z)$, the bias would likely be small when $k = 1$ since it's likely to share a similar label. Meanwhile, when $k \longrightarrow \infty$, the bias doesn't depend on the training points at all which like will restrict it to be higher.

(c) Derive the variance of our model, which is defined as the $Var(h(z))$.

Solution: Let $x_1 \ldots x_k$ be the k closest points.

$$
\begin{aligned}
Var(h(z)) &= Var(\frac{1}{k}\sum_{i=1}^{k} y_i) = \frac{1}{k^2}\sum_{i=1}^{k} Var(f(x_i)+\varepsilon) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}(Var(f(x_i))+Var(\varepsilon)) = \frac{1}{k^2}\sum_{i=1}^{k}(Var(\varepsilon)) \\
&= \frac{1}{k^2}\sum_{i=1}^{k}(\sigma^2) = \frac{1}{k^2}k\sigma^2 = \frac{\sigma^2}{k}
\end{aligned}
$$

(d) What happens to the variance when $k \longrightarrow \infty$? How about when $k = 1$?

Solution:

Variance goes to 0 when $k \longrightarrow \infty$, and is maximized at $k = 1$.

3 MLE, MAP, and Lasso

Assume a set of points $x_1, \dots, x_n \in \mathbb{R}^d$, an unknown parameter vector $\theta^* \in \mathbb{R}^d$, and observations $y_1, \dots, y_n \in \mathbb{R}$ generated by

$$y_i = x_i^\top \theta^* + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some arbitrary σ^2. Note that this can equivalently be written as

$$y_i \sim \mathcal{N}(x_i^\top \theta^*, \sigma^2)$$

(a) Show that performing maximum likelihood estimation under these modeling assumptions is equivalent to solving the unconstrained least squares problem. That is, show that you can formulate the optimization problem as

$$\hat{\theta} = \arg\min_{\theta} \alpha \|X\theta - Y\|_2^2 \tag{1}$$

for $\alpha > 0$, $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$.

Solution:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} p(y_1, \dots, y_n | \theta) \\
&= \arg\max_{\theta} \prod_{i=1}^{n} p(y_i | \theta) \\
&= \arg\max_{\theta} \prod_{i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i^\top \theta - y_i\right)^2}{2\sigma^2} \right) \\
&= \arg\min_{\theta} \sum_{i} \frac{1}{2\sigma^2} \left(x_i^\top \theta - y_i \right)^2 \\
&= \arg\min_{\theta} \|X\theta - Y\|_2^2
\end{aligned}$$

(b) Now assume that θ_i^* is drawn from a distribution with probability density function $p(\theta_i^*) \propto e^{-|\theta_i^*|/t}$ where $t > 0$ is a constant. Show that performing maximum a posteriori estimation is equivalent to solving the l-1 regularized least squares problem. That is, show that you can formulate the optimization problem as

$$\hat{\theta} = \arg\min_{\theta} \alpha \|X\theta - Y\|_2^2 + \beta \|\theta\|_1 \tag{2}$$

for $\alpha > 0$, $\beta > 0$, $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$.

Solution:

$$\begin{aligned}
\hat{\theta} &= \arg\max_{\theta} p(\theta | y_1, \ldots, y_n) \\
&= \arg\max_{\theta} p(y_1, \ldots, y_n | \theta) p(\theta) \\
&= \arg\max_{\theta} p(\theta) \prod_{i=1}^{n} p(y_i | \theta) \\
&= \arg\min_{\theta} \sum_{j=1}^{d} \frac{|\theta_i|}{t} + \sum_i \frac{1}{2\sigma^2} \left(x_i^\top \theta - y_i \right)^2 \\
&= \arg\min_{\theta} \frac{1}{2\sigma^2} \|X\theta - Y\|_2^2 + \frac{1}{t} \|\theta\|_1
\end{aligned}$$

(c) Consider the following l-2 regularized regression problem:

$$\hat{\theta} = \arg\min_{\theta} \|X\theta - Y\|_2^2 + \lambda \|\theta\|_2^2 \tag{3}$$

Solve for $\hat{\theta}$ and show that it is a biased estimator.

Solution:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\left(X^\top X + \lambda I_d \right)^{-1} X^\top Y \right] \\
&= \mathbb{E}\left[\left(X^\top X + \lambda I_d \right)^{-1} X^\top (X\theta^* + \varepsilon) \right] \\
&= \left(X^\top X + \lambda I_d \right)^{-1} X^\top X \theta^* \\
&\neq \theta^*
\end{aligned}$$

(d) Consider the optimization problem below that combines l-1 and l-2 regularization with $\gamma \in [0, 1]$:

$$\hat{\theta} = \arg\min_{\theta} \|X\theta - Y\|_2^2 + \lambda \left[\gamma \|\theta\|_2^2 + (1-\gamma) \|\theta\|_1 \right] \tag{4}$$

Show that it can be rewritten as an l-1 regularized problem with augmented versions of X and Y.

Hint: You can modify X to be a specific block matrix.

Solution: Define augmented versions of X and Y as

$$\tilde{X} = \begin{bmatrix} X \\ \delta I_d \end{bmatrix}$$

$$\tilde{Y} = \begin{bmatrix} Y \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{0}$ is a length d vector of zeros. This implies the following

$$\begin{aligned}\|\tilde{X}\theta - \tilde{Y}\|_2^2 &= \left\| \begin{bmatrix} X\theta - Y \\ \delta\theta \end{bmatrix} \right\|_2^2 \\ &= \|X\theta - Y\|_2^2 + \delta^2\|\theta\|_2^2\end{aligned}$$

Adding on $\kappa\|\theta\|_1$ would result in the elastic-net regularization. Therefore, κ and δ should be chosen as $\kappa = \lambda(1-\gamma)$ and $\delta = \sqrt{\lambda\gamma}$. Then the appropriate form can be obtained:

$$\|\tilde{X}\theta + \tilde{Y}\|_2^2 - \lambda(1-\gamma)\|\theta\|_1 = \|X\theta - Y\|_2^2 + \lambda\gamma\|\theta\|_2^2 + \lambda(1-\gamma)\|\theta\|_1$$

4 GLS and the Gauss-Markov Theorem

Suppose we are in the GLS setting where we have a model $Y = Xw + N$ where $N \sim \mathcal{N}(0, \Sigma)$ for some PSD covariance matrix Σ (that is, the error terms could be correlated). Recall that the GLS estimate is $\hat{w}_{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$ and coincides with the MLE when N is Gaussian. In this problem we will show that the GLS estimator is a "best linear unbiased estimator" of w in that it yields the lowest mean squared error $E(\|\hat{w} - w\|_2^2)$ out of all unbiased estimators $\hat{w}$ of w that are linear in y.

(a) Compute $E(\hat{w}_{GLS})$ and $Cov(\hat{w}_{GLS})$. What is the distribution of $\hat{w}$?

Solution: We compute

$$\begin{aligned}
E(\hat{w}_{GLS}) &= E((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}E(Y) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}E(Xw+N) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Xw \quad (N \text{ has mean } 0) \\
&= w
\end{aligned}$$

Hence $\hat{w}_{GLS}$ is an unbiased estimator of w. For the covariance we calculate

$$\begin{aligned}
Cov(\hat{w}_{GLS}) &= Cov((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Cov(Y)((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1})^T \quad (Cov(Ax) = ACov(x)A^T) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} \\
&= (X^T\Sigma^{-1}X)^{-1}
\end{aligned}$$

In the above derivation we used the fact that if A is symmetric and invertible then A^{-1} is symmetric (this follows from the identity $(A^{-1})^T = (A^T)^{-1}$). $\hat{w}_{GLS}$ is a linear transformation of a multivariate Gaussian, and thus is distributed multivariate Gaussian. The parameters for this distribution are the mean and covariance matrix, which must then be the values for $E(\hat{w}_{GLS})$ and $Cov(\hat{w}_{GLS})$ we calculated above. That is, $\hat{w} \sim \mathcal{N}(w, (X^T\Sigma^{-1}X)^{-1})$.

(b) Show that $MSE(\hat{w}) = E(\|w - \hat{w}\|_2^2)$ can be decomposed into the sum of the squared norm of the bias, $\|w - E(\hat{w})\|_2^2$, and the trace of the covariance matrix $\mathrm{Tr}(Cov(\hat{w}))$. Conclude that for unbiased estimators $\hat{w}$ of w, $MSE(\hat{w}) = \mathrm{Tr}(Cov(\hat{w}))$.

Solution: We have

$$\begin{aligned}
E(\|w - \hat{w}\|_2^2) &= E(\|w - E(\hat{w}) + E(\hat{w}) - \hat{w}\|_2^2) \\
&= E(\|w - E(\hat{w})\|_2^2) + E(\|E(\hat{w}) - \hat{w}\|_2^2) + 2E(\langle w - E(\hat{w}), E(\hat{w}) - \hat{w}\rangle)
\end{aligned}$$

The first term is the expectation of a constant, so $E(\|w - E(\hat{w})\|_2^2) = \|w - E(\hat{w})\|_2^2$, which is the squared norm of the bias. The second term is

$$\begin{aligned}
E(\mathrm{Tr}(\|E(\hat{w}) - \hat{w}\|_2^2)) &= E(\mathrm{Tr}((\hat{w} - E(\hat{w}))^T(\hat{w} - E(\hat{w})))) \\
&= E(\mathrm{Tr}((\hat{w} - E(\hat{w}))(\hat{w} - E(\hat{w}))^T)) = \mathrm{Tr}(Cov(\hat{w}))
\end{aligned}$$

since trace and expectation commute. We show the last term is equal to 0 by expanding as

$$2E(wE(\hat{w}) - w\hat{w} - E(\hat{w})^2 + E(\hat{w})\hat{w}) = 2(wE(\hat{w}) - wE(\hat{w}) - E(\hat{w})^2 + E(\hat{w})^2) = 0$$

Thus we have decomposed $MSE(\hat{w}) = \|w - E(\hat{w})\|_2^2 + \mathrm{Tr}(Cov(\hat{w}))$, hence for unbiased estimators the MSE is the trace of the covariance matrix.

(c) In this part of the problem we will prove a version of the Gauss-Markov Theorem for GLS, which states that if $\hat{w}$ is an unbiased estimator of w that is linear in y (that is, $\hat{w} = Cy$ for some C), then $Cov(\hat{w}) - Cov(\hat{w}_{GLS})$ is positive semi-definite.

(a) Set $M = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$ so that $\hat{w}_{GLS} = MY$. If $\hat{w} = (M+D)Y$ where $D \neq 0$ (because if $D = 0$, $\hat{w} = \hat{w}_{GLS}$), show that a necessary and sufficient condition for $\hat{w}$ to be unbiased for every w is the condition $DX = 0$ (hint: take $E(\hat{w})$ and express it as β plus another term).

Solution: We have

$$\begin{aligned}
E(\hat{w}) &= E((M+D)Y) = E((M+D)(Xw+N)) \\
&= E((M+D)Xw) \quad (N \text{ has mean zero}) \\
&= E(MXw + DXw) = w + DXw
\end{aligned}$$

The last line used the fact $E(MXw) = E((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Xw) = w$. So $\hat{w}$ is unbiased for all choices of w iff $DX = 0$.

(b) Show that $Cov(\hat{w}_{GLS}) - Cov(\hat{w})$ is PSD for every such $\hat{w}$ satisfying the conditions for the Gauss-Markov Theorem (hint: take $Cov(\hat{w})$ and express it as $Cov(\hat{w}_{GLS})$ plus another term using the condition found in part (a) - then show that term is PSD).

Solution: We have

$$\begin{aligned}
Cov(\hat{w}) &= Cov((M+D)Y) \\
&= (M+D)\Sigma(M+D)^T
\end{aligned}$$

We have $M\Sigma M^T = Cov(\hat{w}_{GLS})$. We can compute

$$D\Sigma M^T = D\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1} = DX(X^T\Sigma^{-1}X) = 0$$

since $DX = 0$ from part (a). From this we also know $M\Sigma D^T = (D\Sigma M^T)^T = 0^T = 0$. Thus we have the decomposition

$$Cov(\hat{w}) = Cov(\hat{w}_{GLS}) + D\Sigma D^T$$

It is simple to show $D\Sigma D^T$ is PSD. For any v, $vD\Sigma D^T v = (D^T v)^T\Sigma(D^T v) \geq 0$ because Σ is PSD. Thus $Cov(\hat{w}) - Cov(\hat{w}_{GLS}) = D\Sigma D^T$ is PSD.

(c) Does the Gauss-Markov theorem apply when the errors N do not follow a normal distribution?

Solution: Yes, in our proof we had no distributional assumptions on the error term other than that it has mean 0.

(d) Conclude that the GLS estimator minimizes the MSE over all unbiased estimators that are linear in y. In particular, if the covariance matrix of the errors is not a multiple of the identity, GLS does at least as well as OLS.

Solution: From part 2, the MSE of an unbiased estimator $\hat{w}$ of w is the trace of its covariance matrix, $\text{Tr}(Cov(\hat{w}))$. We can compare the MSE of $\hat{w}$ with that of $\hat{w}_{GLS}$:

$$MSE(\hat{w}) - MSE(\hat{w}_{GLS}) = \text{Tr}(Cov(\hat{w})) - \text{Tr}(Cov(\hat{w}_{GLS})) = \text{Tr}(Cov(\hat{w}) - Cov(\hat{w}_{GLS}))$$

By the Gauss-Markov theorem $Cov(\hat{w}) - Cov(\hat{w}_{GLS})$ is PSD, hence all its eigenvalues are non-negative so its trace is non-negative and so $MSE(\hat{w}_{GLS}) \leq MSE(\hat{w})$.