Guerrilla Section #1

Topics

1. General Linear Algebra
2. Bias Variance
3. MLE/MAP & Regularization
4. Weighted (Generalized) Least Squares

General Linear Algebra

Frobenius Norm, SVD, Trace

Frobenius Norm

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

$$\|M\|_F = \sqrt{tr(MM^H)}$$

Frobenius Norm

A norm must satisfy the following properties:

$\|\alpha A\| = |\alpha| \|A\|$ (being *absolutely homogeneous*)

$\|A + B\| \leq \|A\| + \|B\|$ (being *sub-additive* or satisfying the *triangle inequality*)

$\|A\| \geq 0$ (being *positive-valued*)

$\|A\| = 0$ iff $A = 0_{m,n}$ (being *definite*)

SVD

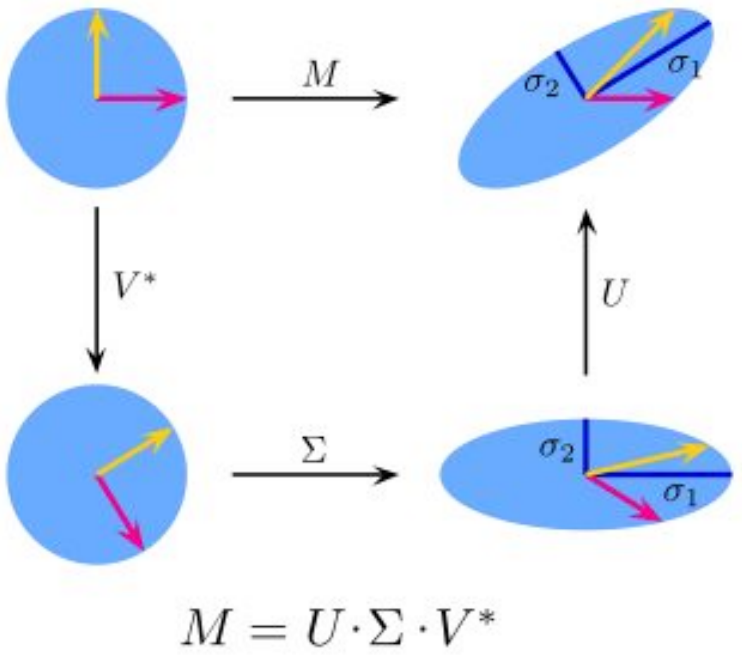
If M is mxn, then U (mxm) and V (nxn) are orthogonal matrices

Σ (mxn) is a diagonal matrix containing singular values in nonincreasing order, σ_i

By spectral theorem:

M^TM (nxn) = VD_vV^T

MM^T (mxm) = UD_uU^T



Trace

Trace is a linear mapping:

$$\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$$
$$\operatorname{tr}(cA) = c\operatorname{tr}(A)$$

The trace of a product can be written as the sum of entry-wise products of elements:

$$\operatorname{tr}\left(X^{\mathrm{T}}Y\right) = \operatorname{tr}\left(XY^{\mathrm{T}}\right) = \operatorname{tr}\left(Y^{\mathrm{T}}X\right) = \operatorname{tr}\left(YX^{\mathrm{T}}\right) = \sum_{i,j} X_{ij}Y_{ij}$$

If A is an nxn matrix with real or complex entries:

$$\operatorname{tr}(A) = \sum_{i} \lambda_i$$

Trace

Cyclic Property

$$\mathrm{tr}(ABCD) = \mathrm{tr}(BCDA) = \mathrm{tr}(CDAB) = \mathrm{tr}(DABC)$$

Arbitrary permutations not allowed...

$$\mathrm{tr}(ABC) \neq \mathrm{tr}(ACB)$$

Unless 3 symmetric matrices (does not apply to more than 3)

(Proof: $\mathrm{tr}(ABC) = \mathrm{tr}(A^{\mathsf{T}} B^{\mathsf{T}} C^{\mathsf{T}}) = \mathrm{tr}(A^{\mathsf{T}}(CB)^{\mathsf{T}}) = \mathrm{tr}((CB)^{\mathsf{T}} A^{\mathsf{T}}) = \mathrm{tr}((ACB)^{\mathsf{T}}) = \mathrm{tr}(ACB)$

Solution 1.1

Solution:

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

$$\|M\|_F^2 = \mathrm{Tr}(MM^T)$$

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

Solution 1.2

Solution:

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

Solution 1.3

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)

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

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

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

Solution 1.3 (Cont.)

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

Bias Variance

The Model

- We assume there is a true function $f(x)$ to describe our regression
- Real world not perfect. For datapoint x, $y(x) = f(x) + N$ where N is random noise
- Generally we model noise N to be normal with 0 mean. e.g. $N \sim \mathbb{N}(0, \Sigma^2)$
- Note that in average case, $\mathbb{E}[y(x)] = \mathbb{E}[f(x) + N] = \mathbb{E}[f(x)] + \mathbb{E}[N] = f(x)$
- Model a hypothesis $= h(x)$, the hypothesis we think is of f's form

Derivation

$$\varepsilon(x;h) = E[(h(x;D) - Y)^2] = E[h(x;D)^2] + E[Y^2] - 2E[h(x;D) \cdot Y]$$

$$= \left(V[h(x;D)] + E[h(x;D)]^2\right) + \left(V[Y] + E[Y]^2\right) - 2E[h(x;D)] \cdot E[Y]$$

$$= \left(E[h(x;D)]^2 - 2E[h(x;D)] \cdot E[Y] + E[Y]^2\right) + V[h(x;D)] + V[Y]$$

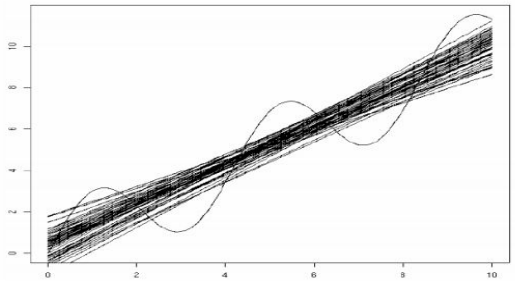
$$= \left(E[h(x;D)] - E[Y]\right)^2 + V[h(x;D)] + V[Y]$$

$$= \underbrace{\left(E[h(x;D)] - f(x)\right)^2}_{bias^2 \text{ of method}} + \underbrace{V[h(x;D)]}_{\text{variance of method}} + \underbrace{V(N)}_{\text{irreducible error}}$$

Qualitative Points

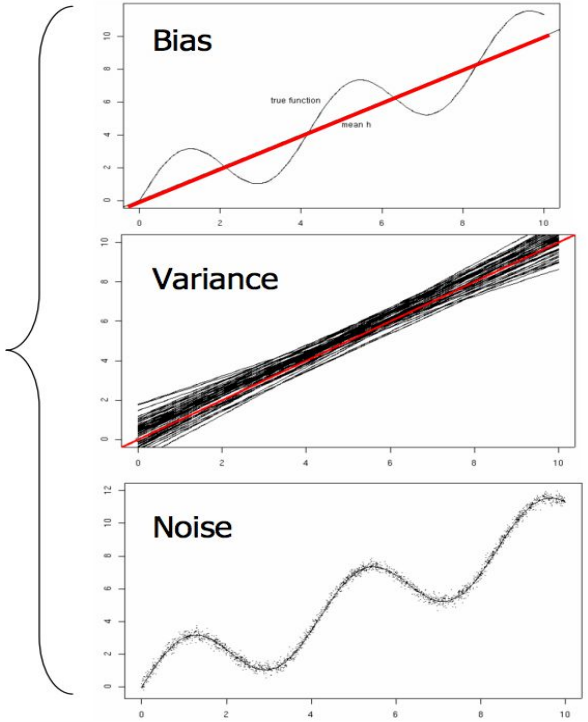
- Bias^2 of method: measures how well the average hypothesis (over all possible training sets) can come close to the true underlying value $f(x)$, for a fixed value of x.
- A low bias means that on average the regressor $h(x)$ accurately estimates $f(x)$.
- Variance of method: Measures the variance of the hypothesis (over all possible training sets), for a fixed value of x.
- A low variance means that the prediction does not change much as the training set varies.
- Irreducible error: This is the error in our model that we cannot control or eliminate, because it is due to errors inherent in our noisy observation Y.

Bias-Variance Visually



50 fits (20 examples each)

=



Problem 2

Solution

1. 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.

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

Solution

2. 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.

Solution

3. 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.

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

Solution

4. 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 $k = 1$.

MLE/MAP & Regularization

Maximum Likelihood Estimation (MLE)

- Assumes uniform priors for all parameter settings

$$f(x_1, x_2, \ldots, x_n \mid \theta) = f(x_1 \mid \theta) \times f(x_2 \mid \theta) \times \cdots \times f(x_n \mid \theta).$$

$$\mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

$$\ln \mathcal{L}(\theta\,;\, x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f(x_i \mid \theta),$$

MLE cont.

$$\theta_{MLE} = \arg\max_{\theta} \log P(X|\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)$$

Maximum A Posteriori

- Assumes non-uniform priors for parameter settings

$$\theta_{MAP} = \arg\max_{\theta} P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log P(X|\theta)P(\theta)$$

$$= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)P(\theta)$$

$$= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta)$$

L-2 Regularization

- Adds a regularization term with weight to restrict size of weights
- Uses 2-norm of weight vector

Find w that minimizes $|Xw - y|^2 + \lambda |w|^2 = J(w)$

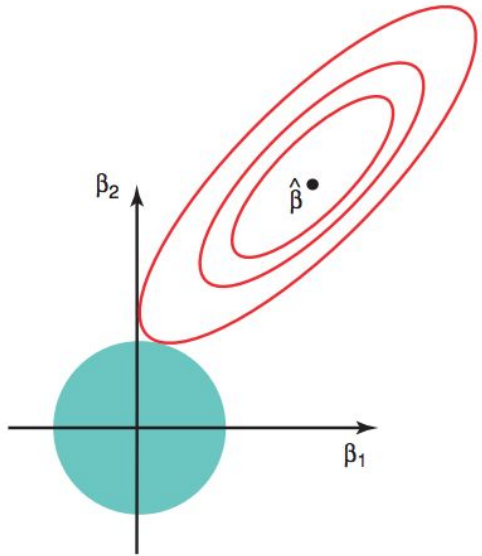


Image obtained from Prof. Shewchuck's CS189 lecture notes from Spring 2017

L-1 Regularization (Lasso)

- Similar to L-2, instead uses 1-norm of weight vector

Find w that minimizes $|Xw - y|^2 + \lambda \|w\|_1 = J(w)$

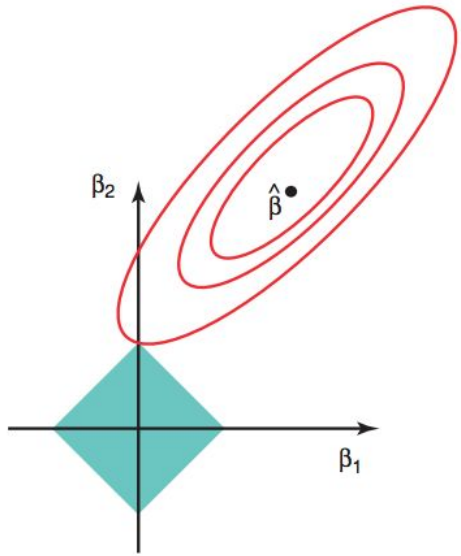
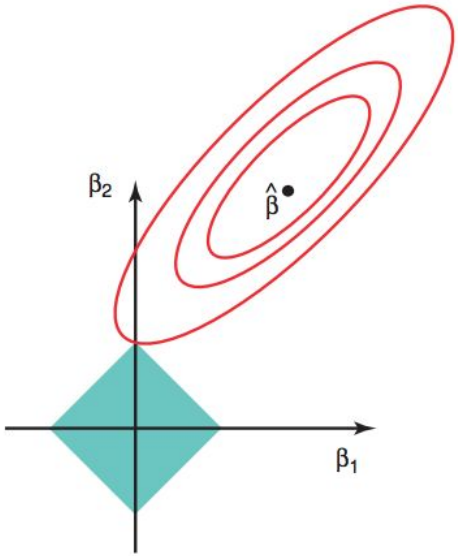
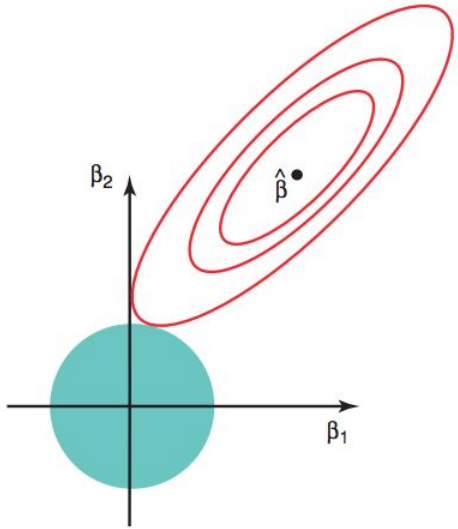


Image obtained from Prof. Shewchuck's CS189 lecture notes from Spring 2017

L-1 vs. L-2





Solution

(a)

$$
\begin{aligned}
\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, p(y_1, \ldots y_n|\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(y_i|\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \prod_{i} \frac{1}{\sqrt{2\pi\alpha^2}} exp(-\frac{(x_i^\top \theta - y_i)^2}{2\sigma^2}) \\
&= \underset{\theta}{\operatorname{argmax}} \sum_{i} \frac{1}{2\sigma^2} (x_i^\top \theta - y_i)^2 \\
&= \underset{\theta}{\operatorname{argmin}} \|X\theta - Y\|_2^2
\end{aligned}
$$

Solution

(b)

$$
\begin{aligned}
\hat{\theta} &= \operatorname*{argmax}_{\theta} p(\theta | y_1, \ldots y_n) \\
&= \operatorname*{argmax}_{\theta} p(y_1, \ldots y_n | \theta) p(\theta) \\
&= \operatorname*{argmax}_{\theta} p(\theta) \prod_{i=1}^{n} p(y_i | \theta) \\
&= \operatorname*{argmin}_{\theta} \sum_{j=1}^{d} \frac{|\theta_i|}{t} + \sum_{i} \frac{1}{2\sigma^2} (x_i^\top \theta - y_i)^2 \\
&= \operatorname*{argmin}_{\theta} \frac{1}{2\sigma^2} \|X\theta - Y\|_2^2 + \frac{1}{t} \|\theta\|_1
\end{aligned}
$$

Solution

(c)

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

Solution

(d) Define augmented versions of X and Y as

$$\tilde{X} = \begin{bmatrix} X \\ \delta I_d \end{bmatrix}$$
$$\tilde{Y} = \begin{bmatrix} Y \\ 0 \end{bmatrix}$$

Where 0 is a length d vector of zeros. This implies the following:

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

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 $\delta = \sqrt{\gamma\lambda}$. Then the appropriate form can be obtained:

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

Weighted/Generalized Least Squares

Probabilistic Model

OLS Model (homoscedastic errors):

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

WLS model (heteroscedastic errors):

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, D)$$

$$D = Diag(\sigma_i^2)$$

GLS model (correlated errors):

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma)$$

MLE for GLS

Maximum likelihood estimation for GLS is equivalent to the following minimization problem:

$$\min_{\beta}(Y - X\beta)^T \Sigma^{-1}(Y - X\beta)$$

In the diagonal covariance matrix case, this is equivalent to weighting the error in the observations by the reciprocal of the variance of the data:

$$\min_{\beta} \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - x_i^T \beta)^2$$

GLS Solution, Limitations

The solution to GLS is given by

$$\hat{\beta}_{GLS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$$

Problem: covariance matrix is not known in advance!

In practice: for weighted least squares with diagonal covariance, can do OLS to obtain estimates of the variance of errors, then WLS using these estimates as weights, repeat until convergence of parameters (this is called iteratively reweighted least squares). For generalized least squares, can make assumptions on the nature of correlation between error terms (for example, noise at one measurement can be modeled as noise at the previous measurement plus random Gaussian).

Question 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.

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

Solution: $E(\hat{w}_{GLS}) = w$ (unbiased), $Cov(\hat{w}_{GLS}) = (X^T \Sigma^{-1} X)^{-1}$. $\hat{w}$ is multivariate normal with the parameters we just found.

Question 2

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.

Question 3

3. 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 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 $D \neq 0$, 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

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

So $\hat{w}$ is unbiased for all choices of w iff $DX = 0$.

Question 3b

(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).

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

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). A similar calculation shows $M\Sigma D^T = 0$. 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 $Cov(\hat{w}) - Cov(\hat{w}_{GLS})$ is PSD.

Question 3c

(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.

Question 4

4. 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, $\mathrm{Tr}(Cov(\hat{w}))$. We can compare the MSE of $\hat{w}$ with that of $\hat{w}_{GLS}$:

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

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 $MSE(\hat{w}_{GLS}) \leq MSE(\hat{w})$.