Introductory Material and Linear Methods

Robert Schmidt

Abstract

The following are notes on the key results from the **Elements of Statistical Learning** text. They were primarily derived from course notes and readings in the Stanford STATS 315: *Modern Applied Statistics* series.

Contents

1 Modeling basics

1.1 Linear model bias-variance decomposition

1) Overview

Conceptually, the bias-variance tradeoff is the problem of how closely one should fit a model to presented data. The tradeoff stipulates that there exists some minimum to the total error such that the model has optimal bias (how different the constructed function is from the true function) and variance (how much, on average, the estimated function differs from its expected value – essentially controlling the closeness of the function fit to the training data).

2) Mathematical specification

Mathematically, this can be derived from the squared-error loss formula, and similarly for other losses.
We consider squared error for illustrative purposes in this case.
Suppose $Y = f(X) + \varepsilon$, for noise ε where $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma_\varepsilon^2$, and errors independent of Y.
Thus, the MSE at a point x_0 is as follows:

$$\begin{aligned} MSE(x_0) &= \mathbb{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0] \\ &= \mathbb{E}[(f(x_0) + \varepsilon - \hat{f}(x_0))^2] \\ &= \mathbb{E}[(\{f(x_0) - \hat{f}(x_0)\} + \varepsilon)^2] \end{aligned}$$

Now, we expand this expression.

$$\begin{aligned} MSE(x_0) &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2 + 2\varepsilon(f(x_0) - \hat{f}(x_0)) + \varepsilon^2] \\ &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + 2\,\mathbb{E}[\varepsilon(f(x_0) - \hat{f}(x_0))] + \mathbb{E}[\varepsilon^2] \quad \text{by linearity of expectation} \end{aligned}$$

First, consider the middle term. By error independence, we can simplify this term significantly.

$$2\,\mathbb{E}[\varepsilon(f(x_0) - \hat{f}(x_0))] = 2\,\mathbb{E}[\varepsilon]\,\mathbb{E}[f(x_0) - \hat{f}(x_0)] = 0 \quad \text{since } \mathbb{E}[\varepsilon] = 0$$

Since the middle term cancels completely, we see that

$$\begin{aligned} MSE(x_0) &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \mathbb{E}[\varepsilon^2] \\ &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \sigma_\varepsilon^2 \\ &= \text{Avg(model error)} + \text{(irreducible error)} \end{aligned}$$

Let us now expand the model error via the bias-variance decomposition.

$$\begin{aligned} \text{Model error} &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] \\ &= \mathbb{E}[(f(x_0) - \mathbb{E}\,\hat{f}(x_0) + \mathbb{E}\,\hat{f}(x_0) - \hat{f}(x_0))^2] \\ &= \mathbb{E}[(f(x_0) - \mathbb{E}\,\hat{f}(x_0))^2 + 2(f(x_0) - \mathbb{E}\,\hat{f}(x_0))(\mathbb{E}\,\hat{f}(x_0) - \hat{f}(x_0)) + (\mathbb{E}\,\hat{f}(x_0) - \hat{f}(x_0))^2] \end{aligned}$$

Again, we will focus on the middle term. If we note that $f(x_0)$ and $\mathbb{E}\,\hat{f}(x_0)$ are not random in regards to the expectation over the data, we can cancel this term:

$$\begin{aligned} 2\,\mathbb{E}[(f(x_0) - \mathbb{E}\,\hat{f}(x_0))(\mathbb{E}\,\hat{f}(x_0) + \hat{f}(x_0))] &= 2(f(x_0) - \mathbb{E}\,\hat{f}(x_0))\,\mathbb{E}[\mathbb{E}\,\hat{f}(x_0) - \hat{f}(x_0)] \\ &= 2(f(x_0) - \mathbb{E}\,\hat{f}(x_0))[\mathbb{E}\,\hat{f}(x_0) - \mathbb{E}\,\hat{f}(x_0)] \\ &= 0 \end{aligned}$$

So, we finally get:

$$\begin{aligned} \text{Model error} &= (\mathbb{E}\,\hat{f}(x_0) - f(x_0))^2 + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}\,\hat{f}(x_0))^2] \\ &= \text{Avg}(\text{bias}^2) + \text{Avg(variance)} \end{aligned}$$

And, returning to the overall average error, we see that:

$$\begin{aligned} MSE(x_0) &= (\mathbb{E}\,\hat{f}(x_0) - f(x_0))^2 + \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}\,\hat{f}(x_0))^2] + \sigma_\varepsilon^2 \\ &= (\text{bias}^2) + \text{(variance)} + \text{(irreducible error)} \\ &= \textbf{bias-variance decomposition} \end{aligned}$$

1.2 K-NN bias-variance decomposition

Consider squared error: $\text{Err}(x_0) = \mathbb{E}[(y - \hat{f}_k(x_0))^2 \mid \mathbf{X} = x_0]$.
Note that $\hat{f}_k$ is the average of the K nearest neighbors:

$$\hat{f}_k(x_0) = \frac{1}{k}\sum_{\ell=1}^{k} y_{(\ell)} = \frac{1}{k}\sum_{\ell=1}^{k}(f(x_{(\ell)}) + \varepsilon_{(\ell)})$$

Assume that f is continuous and $f(x_{(\ell)}) \approx f(x_0) + \frac{1}{k}\sum_{\ell=1}^{k}\varepsilon_{(\ell)})$.
This implies that $\frac{1}{k}\sum_{\ell=1}^{k}(f(x_{(\ell)}) + \varepsilon_{(\ell)}) \approx f(x_0) + \frac{1}{k}\sum_{\ell=1}^{k}\varepsilon_{(\ell)}$.
Hence,

$$\begin{aligned}\text{Var}[\hat{f}_k(x_0)] &= \frac{1}{k^2}\sum_{\ell=1}^{k}\text{Var}(\varepsilon_{(\ell)}) \\ &= \frac{1}{k^2}\cdot k \cdot \sum_{\ell=1}^{k}\text{Var}(\varepsilon) = \frac{1}{k}\sigma_\varepsilon^2\end{aligned}$$

In conclusion:

$$\boxed{\text{Err}(x_0) = \sigma_\varepsilon^2 + \left[f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})\right]^2 + \frac{\sigma_\varepsilon^2}{k}}$$

1.3 Variance of $\hat{\beta}$

1) Assumptions

1. y_i uncorrelated with constant variance σ^2
2. x_i fixed/not random

2) Derivation

Consider the linear model $y = \mathbf{X}\beta + \varepsilon$.

- Let $\mathbf{X}$ be fixed. This implies $\mathbb{E}(y) = \mathbf{X}\beta$.
- $\mathbb{E}(\hat{\beta}) = \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$.
- So,

$$\begin{aligned}\hat{\beta} - \mathbb{E}\,\hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(y - \mathbf{X}\beta) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon\end{aligned}$$

$$\begin{aligned}\implies \text{Var}(\hat{\beta}) &= \mathbb{E}(\hat{\beta} - \mathbb{E}\,\hat{\beta})(\hat{\beta} - \mathbb{E}\,\hat{\beta})^T \\ &= \mathbb{E}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon)(\varepsilon^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathbb{E}(\varepsilon\varepsilon^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\text{Var}(\varepsilon)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \quad \text{since } \mathbb{E}(\varepsilon) = 0\end{aligned}$$

- Recall the assumption that all entries of y are uncorrelated with variance σ^2. Hence,

$$\implies \text{Var}(\epsilon) = \sigma^2\mathbf{I}_N$$

$$\begin{aligned}\implies \text{Var}(\hat{\beta}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\end{aligned}$$

This matches equation (3.8) on page 47 of *ESL*.

1.4 Relationship between t and F statistic for dropping coefficients

Let RSS_0 be the result of the regression $Y \sim X_1 + X_2 + \cdots X_{p-1}$ (for simplicity, I will denote $X_{(p)} = (X_1, \dots, X_p)$), and let RSS_1 be the residual sum of squares from the regression of Y on the whole set of predictors, $Y \sim X_1 + \cdots + X_p$.
For this problem, the F-statistic for dropping the p^{th} predictor is

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} = \frac{RSS_0 - RSS_1}{RSS_1/(N - p - 1)} = \frac{RSS_0 - RSS_1}{\hat{\sigma}^2}$$

For $\hat{\beta}_{(p)}$ as the set of coefficients for the predictor group $X_{(p)}$, we can explicitly define RSS_0:

$$RSS_0 = (Y - X_{(p)}\hat{\beta}_{(p)})^T (Y - X_{(p)}\hat{\beta}_{(p)})$$

To analyze the RSS formulas, we will regress X_p on $X_{(p)}$ to get the residual Z_p.
By regressing $Y \sim X_{(p)} + Z_p$, we get the coefficients $(\hat{\beta}_p, \hat{\beta}_{(p)})$, such that $\hat{\beta}_p = \frac{\langle Y, Z_p \rangle}{\|Z_p\|^2}$.
Thus,

$$RSS_1 = (Y - X_{(p)}\hat{\beta}_{(p)} - Z_p\hat{\beta}_p)^T (Y - X_{(p)}\hat{\beta}_{(p)} - Z_p\hat{\beta}_p)$$

We can now find an expression for the numerator of the F-statistic:

$$\begin{aligned}
RSS_0 - RSS_1 &= 2(Y - X_{(p)}\hat{\beta}_{(p)})^T Z_p \hat{\beta}_p - \|Z_p\|^2 \hat{\beta}_p \\
&= 2Y^T Z_p \hat{\beta}_p - 2X_{(p)}^T Z_p \hat{\beta}_p - \|Z_p\|^2 \hat{\beta}_p \\
&= 2Y^T Z_p \hat{\beta}_p - \|Z_p\|^2 \hat{\beta}_p \text{ since } Z_p, \hat{\beta}_p \text{ are orthogonal} \\
&= \frac{\langle Y, Z_p \rangle^2}{\|Z_p\|^2} = \hat{\beta}_p^2 \|Z_p\|^2
\end{aligned}$$

In summary, we have found that $F = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2}\|Z_p\|^2$.
From ESL, we see that $t = \frac{\hat{\beta}_p}{\hat{\sigma}\sqrt{v_p}}$, so $t^2 = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2 v_p}$. Thus, to prove that $t^2 = F$, we need to focus on v_p.
By the definition of v_p, we know that $v_p = \left[(\mathbf{X}^T\mathbf{X})^{-1}\right]_{pp}$, where $\mathbf{X} = \begin{bmatrix} X_{(p)} & X_p \end{bmatrix}$. So, we see that

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} X_{(p)}^T X_{(p)} & X_{(p)}^T X_p \\ X_p^T X_{(p)} & X_p^T X_p \end{bmatrix}$$

To invert this matrix, we can employ Schur's complement; for $\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}$, the entry we seek in $\mathbf{M}^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$ is of the form $(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}$. First, we will find $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}$:

$$\begin{aligned}
\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} &= X_p^T X_p - (X_p^T X_{(p)})(X_{(p)}^T X_{(p)})^{-1} X_{(p)}^T X_p \\
&= \|X_p - X_p^T (X_{(p)}^T X_{(p)})^{-1} X_p\|^2 \\
&= \|X_p - \frac{\langle X_p, X_{(p)} \rangle}{\|X_{(p)}\|^2}\|^2
\end{aligned}$$

Recall that Z_p is the residual from regressing X_p on $X_{(p)}$, so it is of the form $Z_p = X_p - \frac{\langle X_p, X_{(p)} \rangle}{\|X_{(p)}\|^2}$, which is the expression in the norm above.
Hence, $\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} = \|Z_p\|^2 \implies v_p = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} = (\|Z_p\|^2)^{-1}$.
Plugging this expression into t:

$$t^2 = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2 v_p} = \frac{\hat{\beta}_p^2}{\hat{\sigma}^2}\|Z_p\|^2 = F$$

In conclusion, we see that, when dropping the last of p variables, $\boxed{t^2 = F}$.

2 Ridge regression

2.1 Derivation using the SVD

As derived in ESL, $\hat{\beta}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$.
We will re-write this estimate in terms of the SVD of $\mathbf{X} = \mathbf{UDV}^T$.
Here, I use the convention that:

$$\begin{cases} \mathbf{U} \text{ is } n \times p \text{ orthonormal} \\ \mathbf{D} \text{ is } p \times p \text{ diagonal with singular values } d_i \text{ on the diagonal} \\ \mathbf{V} \text{ is } p \times p \text{ orthogonal} \end{cases}$$

$$\begin{aligned} \hat{\beta}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\ &= (\mathbf{VDU}^T\mathbf{UDV}^T + \lambda\mathbf{I})^{-1}\mathbf{VDU}^T\mathbf{y} \\ &= (\mathbf{VD}^2\mathbf{V}^T + \lambda\mathbf{VV}^T)^{-1}\mathbf{VDU}^T\mathbf{y} \;\text{ since } \mathbf{U}^T\mathbf{U} = \mathbf{I}_{p\times p} = \mathbf{VV}^T \\ &= (\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}^T)^{-1}\mathbf{VDU}^T\mathbf{y} \\ &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^{-1}\mathbf{VDU}^T\mathbf{y} \;\text{ since } (\mathbf{V}^T)^{-1} = \mathbf{V} \text{ given that } \mathbf{V} \text{ is orthogonal} \\ &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{y} \end{aligned}$$

Hence, in terms of the SVD, $\boxed{\hat{\beta}(\lambda) = \mathbf{V}(\mathbf{D^2} + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{y}}$.

2.2 Eigenvalue interpretation

Consider $\hat{\beta}^{\text{ridge}}$ derived from the SVD:

$$\hat{\beta} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{y}$$

We can use this estimate to find $\hat{y}^{\text{ridge}}$:

$$\begin{aligned} \hat{y}^{\text{ridge}} &= \mathbf{X}\hat{\beta}^{\text{ridge}} \\ &= (\mathbf{UDV}^T)\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{y} \;\text{ by the SVD of } \mathbf{X} \\ &= \mathbf{UD}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^T\mathbf{y} \end{aligned}$$

Consider further the innermost matrix calculation $\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}$. This will be a diagonal matrix with elements $\frac{d_j^2}{d_j^2+\lambda}$:

$$(\mathbf{D^2} + \lambda\mathbf{I})^{-1}\mathbf{D} = \begin{bmatrix} \frac{d_1}{d_1^2+\lambda} & & & \\ & \ddots & & \\ & & \frac{d_j}{d_j^2+\lambda} & \\ & & & \ddots \end{bmatrix}$$

Next, we note that $\mathbf{U}^T y$ represents the coordinates of y in the basis spanned by the p columns of $\mathbf{U}$.
So, summing over the columns, we see that:

$$\begin{aligned} \hat{y}^{\text{ridge}} &= \mathbf{X}\hat{\beta}^{\text{ridge}} \\ &= \sum_{j=1}^{p} u_j \left(\frac{d_j^2}{d_j^2 + \lambda} \right) u_j^T y \\ &= \sum_{j=1}^{p} u_j \left(\frac{d_j^2}{d_j^2 + \lambda} \right) \langle u_j,\, y \rangle \end{aligned}$$

Thus, we see that ridge shrinks and regularizes according to the eigenvalues. Specifically, ridge has the greatest shrinkage effect on coefficients with small d_j corresponding to the directions in the column space of $\mathbf{X}$ having small variance.

2.3 Relationship between λ and the coefficients

Again, recall that the ridge fit vector is of the form

$$\hat{\beta}(\lambda) = \mathbf{V}(\mathbf{D^2} + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}$$

$$\text{where: } \begin{cases} \mathbf{U} \text{ is } n \times p \text{ orthonormal} \\ \mathbf{D} \text{ is } p \times p \text{ diagonal with singular values } d_i \text{ on the diagonal} \\ \mathbf{V} \text{ is } p \times p \text{ orthogonal} \end{cases}$$

Using this decomposition, we now consider $\|\hat{\beta}_\lambda\|_2^2 = \|\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}\|_2^2$.
First, note that this expression is of the form $\|\hat{\beta}_\lambda\|^2 = \|\mathbf{V}\theta\|^2$, for $\theta = (\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}$.
So, our first simplification is that

$$\begin{aligned} \|\hat{\beta}_\lambda\|^2 = \|\mathbf{V}\theta\|^2 &= (\mathbf{V}\theta)^T(\mathbf{V}\theta) \\ &= \theta^T\mathbf{V}^T\mathbf{V}\theta \\ &= \theta^T\theta = \|\theta\|^2 \text{ since } \mathbf{V} \text{ is orthogonal} \end{aligned}$$

Continuing with the expansion of the norm expression, $\|\hat{\beta}_\lambda\|^2 = \|\theta\|^2 = \|(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}\|^2$.
Again taking particular note of the expressions involving $\mathbf{D}$, we recall that:

$$(\mathbf{D^2} + \lambda\mathbf{I})^{-1}\mathbf{D} = \begin{bmatrix} \frac{d_1}{d_1^2+\lambda} & & & \\ & \ddots & & \\ & & \frac{d_j}{d_j^2+\lambda} & \\ & & & \ddots \end{bmatrix}$$

So, the j^{th} component of θ is

$$\left[(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y}\right]_j = \frac{d_j}{d_j^2 + \lambda}\langle\mathbf{U}, \mathbf{y}\rangle_j$$

Hence,

$$\|\hat{\beta}_\lambda\|^2 = \sum_j \left(\frac{d_j}{d_j^2 + \lambda}\right)^2 \left(\langle\mathbf{U}, \mathbf{y}\rangle_j\right)^2$$

We therefore see that as $\lambda \downarrow$, $\boxed{\|\hat{\beta}_\lambda\|^2 \uparrow}$

2.4 Bayesian theory: ridge estimate is the mean of the posterior

Assume that $\mathbf{Y} \mid \beta \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ and $\beta \sim N(0, \tau\mathbf{I})$. Thus:

$$f(\mathbf{Y} \mid \beta) \propto \exp\left(- \frac{\|\mathbf{Y} - \mathbf{X}\beta\|^2}{2\sigma^2} \right) \quad \text{and} \quad f(\beta) \propto \exp\left(- \frac{\|\beta\|^2}{2\tau^2} \right)$$

Note that $f(\beta \mid \mathbf{Y}) = \frac{f(\beta,\mathbf{Y})}{f(\mathbf{Y})} = \frac{f(\beta)f(\mathbf{Y}|\beta)}{f(\mathbf{Y})} \propto f(\beta)f(\mathbf{Y} \mid \beta)$ since $f(\mathbf{Y})$ is a constant in terms of the function.
So, plugging in the expressions we found for these functions,

$$f(\beta \mid \mathbf{Y}) \propto \exp\left(- \frac{\|\mathbf{Y} - \mathbf{X}\beta\|^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2} \right) = \exp\left(- \left(\frac{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2} + \frac{\beta^T\beta}{2\tau^2} \right) \right)$$

We will focus our simplification efforts on the function inside the exponential.

$$\begin{aligned} \frac{(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)}{2\sigma^2} + \frac{\beta^T\beta}{2\tau^2} &= \frac{\mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta}{2\sigma^2} + \frac{\beta^T\beta}{2\tau^2} \\ &= \beta^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2} \right)\beta - \left(\frac{\mathbf{Y}^T\mathbf{X}}{\sigma^2} \right)\beta + c \end{aligned}$$

This expression is quadratic in β; we will factor it into the form $(\beta - \mu)^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2}\right)(\beta - \mu) + c$.
Such a factorization would show that $f(\beta \mid \mathbf{Y}) \propto \exp\left((\beta - \mu)^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2}\right)(\beta - \mu)\right)$, which would imply that $E[f(\beta \mid \mathbf{Y})] = \mu$ since $f(\beta \mid \mathbf{Y})$ would be distributed multivariate Gaussian.
Since $(\beta - \mu)^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2}\right)(\beta - \mu) + c = \beta^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2}\right)\beta - \left(\frac{\mathbf{Y}^T\mathbf{X}}{\sigma^2}\right)\beta + c'$,
we need $-2\mu^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2}\right)\beta = -\left(\frac{\mathbf{Y}^T\mathbf{X}}{\sigma^2}\right)\beta$.
Focusing on this expression to solve for μ:

$$-2\mu^T\left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2} \right)\beta = -\left(\frac{\mathbf{Y}^T\mathbf{X}}{\sigma^2} \right)\beta$$

$$\implies \mu = \left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2} \right)^{-1} \left(\frac{\mathbf{Y}^T\mathbf{X}}{2\sigma^2} \right)^T$$

$$\implies \mu = \left(\frac{\mathbf{X}^T\mathbf{X}}{2\sigma^2} + \frac{\mathbf{I}}{2\tau^2} \right)^{-1} \left(\frac{\mathbf{X}^T\mathbf{Y}}{2\sigma^2} \right) = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I} \right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Thus, $\mu = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ is the mean of the posterior distribution, $f(\beta \mid \mathbf{Y})$.
Recall that the ridge regression solution is $\hat{\beta}(\lambda) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$. This solution is identical to our result with the stipulation that $\boxed{\lambda = \sigma^2/\tau^2}$. Hence, the ridge estimate is indeed the mean of the posterior distribution.

3 Degrees of freedom

3.1 Ridge df

3.2 Lasso df

4 Methods using derived input directions

4.1 Principal component regression up to j coefficients

4.2 PCR connection to OLS/ridge

4.3 PCR vs. ridge

5 Zero-residual fits

5.1 There are infinitely many OLS solutions with zero residuals

5.2 Ridge solution has minimum norm among all zero-residual solutions as $\lambda \to 0$