# Tutorial Linear Regression

Suggested Reading:
Murphy, chapter 7
OR
Bishop, chapter 3

# 1 Linear regression

**Problem 1:** Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$$

takes any vector and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to show that the least square solution for linear regression corresponds to an orthogonal projection of the vector $\boldsymbol{T}$ (denoted by $\boldsymbol{Z}$ in class!) onto the manifold $S$ as shown in Figure 1. There, the subspace $S$ is spanned by the basis functions $\phi_j(\boldsymbol{x})$ in which each basis function is viewed as a vector $\boldsymbol{\varphi}_j$ of length $N$ with elements $\phi_j(\boldsymbol{x}_n)$. (Hint: You might want consider what $\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$ resembles, e.g. how does it relate to the maximum likelihood solution for linear regression.)

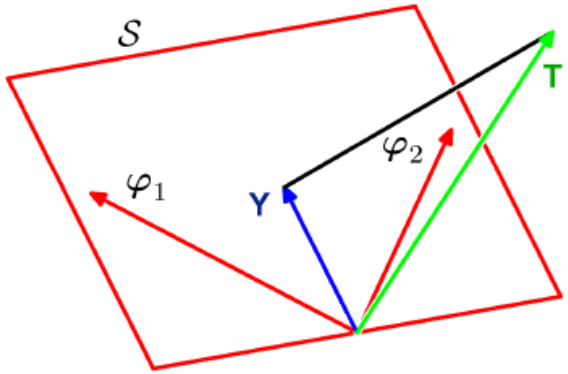


Figure 1: The projection property of $\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$.

If we set $\boldsymbol{\Theta} = \boldsymbol{I}$ in the section on weighted linear regression, we see that for the standard maximum likelihood problem we want to solve $\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w} = \boldsymbol{\Phi}^T\boldsymbol{z}$ which is equivalent to $\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{z}) = 0$. For the solution $\boldsymbol{w_{MLE}}$ one can deduce from the last equation that (i) $\boldsymbol{\Phi w_{ML}}$ is a vector in $S$ and (ii) $\boldsymbol{\Phi w_{ML}} - \boldsymbol{z}$ (the *error* vector between $\boldsymbol{z}$ and $\boldsymbol{\Phi w_{ML}}$) is orthogonal to $S$. Thus $\boldsymbol{\Phi w_{MLE}}$ is the orthogonal projection of $\mathbf{T}$ on $S$. is $\boldsymbol{\Phi}$, thus the whole matrix resembles a projection into a space spanned by the columns of $\boldsymbol{\Phi}$. To show that the least square solution corresponds to an *orthogonal* projection consider the dot product of the projection and the difference vector between $\boldsymbol{t}$ and the projection, that is

$$(t - \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T t)^T(\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T t)$$

which is

$$t^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T t - t^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T t = 0$$

# 2 Ridge regression

**Problem 2:** Using singular value decomposition of the design matrix $\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$ show that the output on the training set fitted with the ridge regression solution $\hat{\boldsymbol{w}}^{ridge}$ can be written as

$$\sum_j \left( \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^T \right) \boldsymbol{z}$$

where $\boldsymbol{u}_j$ are the columns of $\boldsymbol{U}$, $d_j$ the elements of $\boldsymbol{D}$ and $\lambda$ the cost factor of the $\ell 2$ regularization. What is the interpretation of this formula?

Based on the SVD of $\boldsymbol{\Phi}$, we can write (the trick is here to rewrite $\lambda \boldsymbol{I}$ as $\lambda \boldsymbol{V}\boldsymbol{V}^T$ and factor matrices out (and remembering that $\boldsymbol{A}\boldsymbol{B}^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$)):

$$\hat{\boldsymbol{w}}^{ridge} = \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{z}$$

Then

$$\boldsymbol{\Phi}\hat{\boldsymbol{w}}^{ridge} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{z} = \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{z}$$

First observation, $\boldsymbol{D}^2 + \lambda \boldsymbol{I}$ is a diagonal matrix, with $d_j^2 + \lambda$ on the diagonal, $d_j$ are the singular values of $\boldsymbol{\Phi}$. Therefore, its inverse is again a diagonal, with $1/d_j^2 + \lambda$ on the diagonal. And therefore $\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1}\boldsymbol{D}$ is also a diagonal matrix (product of diagonal matrices), with $\frac{d_j^2}{d_j^2+\lambda}$ on the diagonal. This matrix gets multiplied from the right to $\boldsymbol{U}$, i.e. it only scales the columns. Finally the product of two matrices of the form $\boldsymbol{A}\boldsymbol{B}^T$ can be written as the sum of the outer product of the respective columns of $\boldsymbol{A}$ and $\boldsymbol{B}$.

Now let's put the formula in words. First, $\boldsymbol{u}_j^T\boldsymbol{y}$ computes the representation of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$, and then reconstructs $\boldsymbol{y}$ in this basis, however with the coordinates *shrunk* ($\lambda > 0$ and thus $\frac{d_j^2}{d_j^2+\lambda} < 1$). A greater amount of shrinkage is applied to the coordinates with smaller singular values. (What does a small singular value mean? We will later discuss that the SVD of $\boldsymbol{\Phi}$ is another way of expressing the *principal components* of the variables in $\boldsymbol{\Phi}$. These are directions in the space spanned by the training examples in which the training data varies, small singular values are directions in which the training data varies very little. Hence, ridge regression shrinks those directions most. The implicit assumption (or justification for this behaviour) is that the output will vary most with those directions that vary most.)

# 3 Multi-output linear regression

**Problem 3:** In class, we only considered functions of the form $f : \mathbb{R}^n \to \mathbb{R}$. What about the general case of $f : \mathbb{R}^n \to \mathbb{R}^m$? For linear regression with multiple outputs, write down the loglikelihood formulation and derive the MLE of the parameters.

The observation $\boldsymbol{z}_i$ is a vector with $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{\Sigma})$, $\boldsymbol{W} \in R^{m \times n}$, $\boldsymbol{\Sigma} \in R^{m \times m}$, covariance $\boldsymbol{\Sigma}$ is known and fixed for all possible observations. For $n$ i.i.d observed pairs $(\boldsymbol{x}_i, \boldsymbol{z}_i)$, the likelihood is $\prod_i \mathcal{N}(\boldsymbol{W}\boldsymbol{x}_i, \boldsymbol{\Sigma})$, and thus the negative log-likelihood is const $+ \frac{1}{2}\sum_i (\boldsymbol{z}_i - \boldsymbol{W}\boldsymbol{x}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}_i - \boldsymbol{W}\boldsymbol{x}_i)$ Let $\boldsymbol{L}\boldsymbol{L}^T = \boldsymbol{\Sigma}$ (the cholesky decomposition of $\boldsymbol{\Sigma}$), so $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^{-1}\boldsymbol{L}^{-T}$. Using this decomposition write $\boldsymbol{L}^{-T}(\boldsymbol{z}_i - \boldsymbol{W}\boldsymbol{x}_i) = (\boldsymbol{L}^{-T}\boldsymbol{z}_i - \boldsymbol{L}^{-T}\boldsymbol{W}\boldsymbol{x}_i) = (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)$. Using this transformation, the negative log-likelihood now becomes const $+ \frac{1}{2}\sum_i (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)^T (\hat{\boldsymbol{z}}_i - \hat{\boldsymbol{W}}\boldsymbol{x}_i)$. Using similar reasoning to the lecture (or Problem 2 above), we can rewrite this as $(\boldsymbol{\Phi}\hat{\boldsymbol{W}} - \hat{\boldsymbol{Z}})^T (\boldsymbol{\Phi}\hat{\boldsymbol{W}} - \hat{\boldsymbol{Z}})$. Note that $\hat{\boldsymbol{Z}}$ is a matrix that has the vectors $\hat{\boldsymbol{z}}_i$ as its rows. Matrix calculus (derivative of the negative log-likelihood with respect to $\hat{\boldsymbol{W}}$) then gives us $\hat{\boldsymbol{W}}_{MLE} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\hat{\boldsymbol{Z}}$. So these are $m$ single least square problems for every *column* of $\hat{\boldsymbol{Z}}$. Finally, transforming back $\hat{\boldsymbol{W}}_{MLE}$ gives $\boldsymbol{W}_{MLE} = \boldsymbol{L}^T\hat{\boldsymbol{W}}_{MLE}$.

# 4 Bayesian Linear Regression

**Problem 4:** We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Prove the following matrix identity

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}$$

and, using it, show that the uncertainty $\sigma_N^2(\boldsymbol{x})$ associated with the bayesian linear regression function given by

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}) \tag{1}$$

(where $\boldsymbol{S}_N$ is the covariance of the posterior $p(\boldsymbol{w}|\mathcal{D})$. (see e.g. equations 3.53 and 3.54 in Bishop)) satisfies

$$\sigma_{N+1}^2(\boldsymbol{x}) \leq \sigma_N^2(\boldsymbol{x}) \tag{2}$$

You may want to use

$$\boldsymbol{\Phi}_{N+1}^T \boldsymbol{\Phi}_{N+1} = \boldsymbol{\Phi}_N^T \boldsymbol{\Phi}_N + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T$$

Starting from

$$(\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T)^{-1} = \boldsymbol{M}^{-1} - \frac{(\boldsymbol{M}^{-1}\boldsymbol{v})(\boldsymbol{v}^T\boldsymbol{M}^{-1})}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}$$

multiply

$$\boldsymbol{M} + \boldsymbol{v}\boldsymbol{v}^T$$

on both sides, giving

$$\boldsymbol{I} = \boldsymbol{I} + \boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T - \left(\frac{\boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T}{1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}}\right)$$

$\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}$ is a scalar:

$$\boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T + \boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T = (1 + \boldsymbol{v}^T\boldsymbol{M}^{-1}\boldsymbol{v})\boldsymbol{M}^{-1}\boldsymbol{v}\boldsymbol{v}^T$$

To prove eq. (1), we first need the definition of $\sigma_N^2$ again and also that of $\boldsymbol{S}_N$ ($\boldsymbol{\Phi}_N$ denotes the design matrix containing the first $N$ observations).

$$\begin{aligned} \sigma_N^2(\boldsymbol{x}) &= \frac{1}{\beta} + \phi(\boldsymbol{x})^T S_N \phi(\boldsymbol{x}) \\ \boldsymbol{S}_N^{-1} &= \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N \end{aligned}$$

We now use the following identity:

$$\boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N = \sum_{k=1}^{N} \phi(\boldsymbol{x}_k)\phi(\boldsymbol{x}_k)^T$$

so we can write

$$\boldsymbol{\Phi}_{N+1}^T\boldsymbol{\Phi}_{N+1} = \boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T$$

and hence

$$\boldsymbol{S}_{N+1}^{-1} = \alpha\boldsymbol{I} + \beta\boldsymbol{\Phi}_{N+1}^T\boldsymbol{\Phi}_{N+1} = \alpha\boldsymbol{I} + \boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T = \boldsymbol{S}_N^{-1} + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T$$

Substituting this into $\sigma^2_{N+1}(\boldsymbol{x})$ and using the above matrix identity we get:

$$\begin{aligned}
\sigma^2_{N+1}(\boldsymbol{x}) &= \frac{1}{\beta} + \phi(\boldsymbol{x})^T \boldsymbol{S}_{N+1} \phi(\boldsymbol{x}) \\
&= \frac{1}{\beta} + \phi(\boldsymbol{x})^T (\boldsymbol{S}_N^{-1} + \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T)^{-1} \phi(\boldsymbol{x}) \\
&= \frac{1}{\beta} + \phi(\boldsymbol{x})^T \left( \boldsymbol{S}_N - \frac{\boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T \boldsymbol{S}_N}{1 + \phi(\boldsymbol{x}_{N+1})^T \boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1})} \right) \phi(\boldsymbol{x}) \\
&= \frac{1}{\beta} + \phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}) - \frac{\phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1})\phi(\boldsymbol{x}_{N+1})^T \boldsymbol{S}_N \phi(\boldsymbol{x})}{1 + \phi(\boldsymbol{x}_{N+1})^T \boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1})}
\end{aligned}$$

The nominator of the last fraction $(\phi(\boldsymbol{x})^T \boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1}))^2 \geq 0$. Also (with $\alpha, \beta \geq 0$)

$$\phi(\boldsymbol{x}_{N+1})^T \boldsymbol{S}_N \phi(\boldsymbol{x}_{N+1}) = \phi(\boldsymbol{x}_{N+1})^T \alpha \boldsymbol{I} \phi(\boldsymbol{x}_{N+1}) + \phi(\boldsymbol{x}_{N+1})^T \beta \boldsymbol{\Phi}_N^T \boldsymbol{\Phi}_N \phi(\boldsymbol{x}_{N+1}) \geq 0$$

so we have proven eq. (1).

**Problem 5:** We know that the posterior for a linear regression algorithm with a likelihood defined by $p(Z \mid W, \beta) = \prod_{n=1}^{N} \mathcal{N}(Z_n \mid W^T \Phi(X_n), \beta^{-1})$ and prior given by $p(w) = \mathcal{N}(W \mid M_0, \mathbf{S_0})$ is

$$p(W \mid Z) = \mathcal{N}(W \mid M_N, \mathbf{S}_N)$$

where

$$M_N = \mathbf{S}_N(S_0^{-1} M_0 + \beta \Phi^T Z)$$
$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi$$

Let's assume $\beta$ is a known constant. Verify that this is the form of the posterior we would derive.

From Bayes' theorem we have

$$p(W \mid Z) \propto p(Z \mid W)p(W),$$

where the factors on the r.h.s. are given by the equations listed above. Writing this our in full, we get

$$p(W \mid Z) \propto [\prod_{n=1}^{N} \mathcal{N}(Z_n \mid W^T \Phi(X_n), \beta^{-1})]\mathcal{N}(w \mid M_0, \mathbf{S}_0)$$

$$\propto \exp(\frac{-\beta}{2}(Z - \mathbf{\Phi}W)^T(Z - \mathbf{\Phi}W)) \exp(\frac{-1}{2}(W - M_0)^T\mathbf{S}_0^{-1}(W - M_0))$$

$$= \exp\left(\frac{-1}{2}(W^T(\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}\mathbf{\Phi}^T)W - \beta W^T\mathbf{\Phi}^T Z + \beta Z^T Z + M_0^T\mathbf{S}_0^{-1}W - W^T\mathbf{S}_0^{-1}M_0 + M_0^T\mathbf{S}_0^{-1}M_0)\right)$$

$$= \exp\left(\frac{-1}{2}(W^T(\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi})W - (\mathbf{S}_0^{-1}M_0 + \beta\mathbf{\Phi}^T Z)^T W - W^T(\mathbf{S}_0^{-1}M_0 + \beta\mathbf{\Phi}^T Z) + \beta Z^T Z + M_0^T\mathbf{S}_0^{-1}M_0)\right)$$

$$= \exp\left(\frac{-1}{2}(w - M_N)^T\mathbf{S}_N^{-1}(W - M_N)\right) \exp\left(\frac{-1}{2}(\beta Z^T Z + M_0^T\mathbf{S}_0^{-1}M_0 - M_N^T\mathbf{S}_N^{-1}M_N)\right)$$

where we used the definitions for $M_N$ and $\mathbf{S}_N^{-1}$ from above when completing the square in the last step. The first exponential corresponds to the posterior, unnormalized Gaussian distribution over $W$, while the second exponential is independent of $W$ and hence can be absorbed into the normalization factor.

# 5 Online Learning

**Problem 6:** Suppose we are using a linear basis function model where the posterior distribution is given by $p(W \mid Z) = \mathcal{N}(W \mid M_N, \mathbf{S}_N)$ and we have already observed N data points. That means that this posterior can be regarded as the prior for the next observation. By considering an additional data point $(X_{N+1}, z_{N+1})$, and by completing the square in the exponential, show that the resulting posterior distribution is again given by the posterior mentioned above, but with $\mathbf{S}_N$ replaced by $\mathbf{S}_{N+1}$ and $M_N$ replaced by $M_{N+1}$.

Combining the prior

$$p(W) = \mathcal{N}(W \mid M_N, \mathbf{S}_N)$$

and the likelihood

$$p(Z_{N+1} \mid X_{N+1}, W) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \exp\left(\frac{-\beta}{2}(Z_{N+1} - W^T\Phi_{N+1})^2\right)$$

where $\Phi_{N+1} = \Phi(X_{N+1})$, we obtain a posterior of the form

$$p(W \mid Z_{N+1}, X_{N+1}, M_N, \mathbf{S}_N) \propto \exp\left(-\frac{1}{2}(W - M_N)^T\mathbf{S}_N^{-1}(W - M_N) - \frac{1}{2}\beta(Z_{N+1} - W^T\Phi_{N+1})^2\right).$$

We can expand the argument of the exponential, omitting the $-1/2$ factors, as follows

$$(W - M_N)^T\mathbf{S}_N^{-1}(W - M_N) + \beta(Z_{N+1} - W^T\Phi_{N+1})^2$$

$$= W^T\mathbf{S}_N^{-1}W - 2W^T\mathbf{S}_N^{-1}M_N + \beta W^T\Phi_{N+1}^T\Phi_{N+1}W - 2\beta W^T\Phi_{N+1}Z_{N+1} + const$$

$$= W^T(\mathbf{S}_N^{-1} + \beta\Phi_{N+1}\Phi_{N+1}^T)W - 2W^T(\mathbf{S}_N^{-1} + \beta\Phi_{N+1}Z_{N+1}) + const,$$

where const denotes remaining terms independent of $W$. From this we can read off the desired result directly,

$$p(W \mid Z_{N+1}, X_{N+1}, M_N, \mathbf{S}_N) = \mathcal{N}(W \mid M_{N+1}, \mathbf{S}_{N+1}),$$

with

$$\mathbf{S}_{N+1}^{-1} = \mathbf{S}_N^{-1} + \beta\Phi_{N+1}\Phi N + 1^T$$

and

$$M_{N+1} = \mathbf{S}_{N+1}(\mathbf{S}_N^{-1}M_N + \beta\Phi_{N+1}Z_{N+1}).$$