

CS195-5, Lectures 5-6: Derivations and notes  
by Greg Shakhnarovich `gregory@cs`

## Remark on Lecture 4 material

I would like to clarify some potential confusion. As we saw in the lecture, the ML estimate of regression parameters  $\hat{\mathbf{w}}$  is distributed as a Gaussian random variable, with mean  $\mathbf{w}^*$  and covariance  $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ . A surprising implication of this fact is that the distribution of  $\hat{\mathbf{w}}$ , and in particular the amount of uncertainty about  $\mathbf{w}^*$  as measured by  $|\text{Cov}_{\hat{\mathbf{w}}}|$ , does not depend on the training labels  $\mathbf{Y}$ s - only on inputs  $\mathbf{X}$ !

However, there is a subtle but important difference between the actual value of  $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$  and the value we *estimate* based on the data. In particular, in order to estimate this quantity we need an estimate of  $\sigma^2$ . As will be shown in the problem set, the ML estimate of  $\sigma^2$  *does* depend on the training labels. To summarize: the underlying uncertainty about  $\mathbf{w}^*$  inherent in our estimate of  $\hat{\mathbf{w}}$  depends on  $\mathbf{X}$  only. Our *estimate* of this uncertainty depends on both  $\mathbf{X}$  and on  $\mathbf{Y}$ .

## Lecture 5, slide 7

Why do we say that correlation measures *linear* relationship and not just any relationship? Consider the following pair of random variables:

$x$  is uniformly distributed between -1 and 1;

$$y = \sqrt{1 - x^2}.$$

Obviously,  $x$  and  $y$  are highly dependent: given  $x$ ,  $y$  is completely deterministic. However, as it turns out, they are not correlated!

It is a useful exercise to convince yourself that this is true. A good first step is to generate a large sample of values of  $x$  and  $y$  and compute `corrcoef(x, y)`; then do the math, and compute the covariance matrix (from which correlation is obtained trivially). You will need to look up, or derive, the mean and variance of the uniform density, and also recall how to compute expectations of a function of a random variable (for  $y$ ).

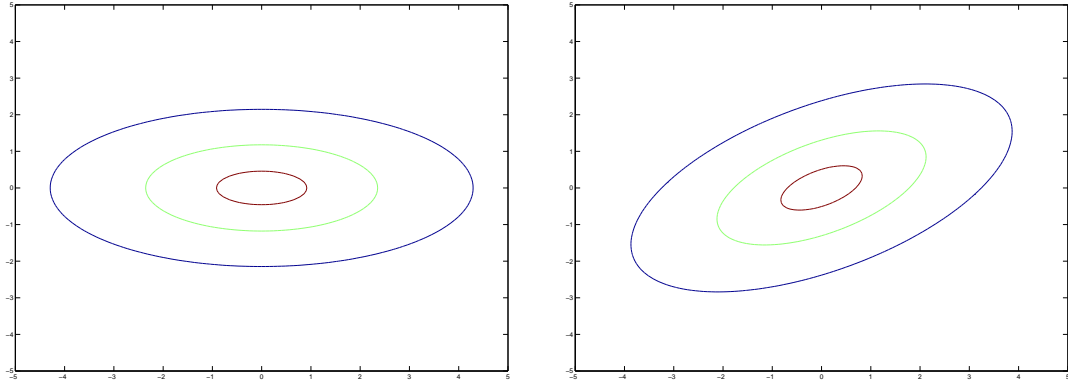


Figure 1: Left:  $\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Lambda})$ . Right:  $\mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma})$ .

## Lecture 5, slide 10

Here is a problem that may make the geometry of Gaussians a bit more intuitive (on a 2D example, of course). Suppose you want to construct a covariance matrix for which the ellipse has the major axis rotated by 30 degrees, and which has variances 4 and 1 along the major and the minor axes, respectively.

The corresponding decomposition of the desired  $\mathbf{\Sigma}$  is as follows. The scaling corresponds to

$$\mathbf{\Lambda} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}.$$

If we just set  $\mathbf{\Sigma} = \mathbf{\Lambda}$  we will get the Gaussian on the left in Fig. 1. To rotate it, we need to multiply the covariance on both sides by the rotation matrix corresponding to rotating the  $x$  axis by 30 degrees. Constant density contours for the resulting Gaussian are plotted on the right in Fig. 1.

## Lecture 6, slide 13

Our goal is to find  $\mathbf{w}$  that maximizes

$$J(\mathbf{w}) = \frac{(\mathbf{w}^T(\mathbf{m}_{+1} - \mathbf{m}_{-1}))^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \quad (1)$$

where  $S_W = N_{+1}\mathbf{S}_{+1} + N_{-1}\mathbf{S}_{-1}$ .

We will denote by  $\mathbf{S}_B$  the *between-class scatter*

$$\mathbf{S}_B \triangleq (\mathbf{m}_{+1} - \mathbf{m}_{-1})(\mathbf{m}_{+1} - \mathbf{m}_{-1})^T \quad (2)$$

so that the numerator in (1) becomes

$$(\mathbf{w}^T(\mathbf{m}_{+1} - \mathbf{m}_{-1}))^2 = \mathbf{w}^T \mathbf{S}_B \mathbf{w}.$$

and consequently

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}.$$

The derivative with respect to  $\mathbf{w}$ :

$$\frac{d}{d\mathbf{w}} J(\mathbf{w}) = \frac{\left(\frac{d}{d\mathbf{w}} \mathbf{w}^T \mathbf{S}_B \mathbf{w}\right) \mathbf{w}^T \mathbf{S}_W \mathbf{w} - \mathbf{w}^T \mathbf{S}_B \mathbf{w} \left(\frac{d}{d\mathbf{w}} \mathbf{w}^T \mathbf{S}_W \mathbf{w}\right)}{(\mathbf{w}^T \mathbf{S}_W \mathbf{w})^2}, \quad (3)$$

very similar to the derivative of  $f(x)/g(x)$  for scalar functions  $f, g$ . Derivative of a quadratic form is given by

$$\frac{d}{dx} \mathbf{x}^T \mathbf{A} \mathbf{x} = 2\mathbf{A} \mathbf{x},$$

so setting (3) to zero we get

$$\frac{d}{d\mathbf{w}} J(\mathbf{w}) = \frac{(2\mathbf{S}_B \mathbf{w}) \mathbf{w}^T \mathbf{S}_W \mathbf{w} - \mathbf{w}^T \mathbf{S}_B \mathbf{w} (2\mathbf{S}_W \mathbf{w})}{(\mathbf{w}^T \mathbf{S}_W \mathbf{w})^2} = 0 \quad (4)$$

Dividing both sides by 2 and multiplying by the scalar  $\mathbf{w}^T \mathbf{S}_W \mathbf{w}$ , we get

$$\frac{\mathbf{S}_B \mathbf{w} \mathbf{w}^T \mathbf{S}_W \mathbf{w} - \mathbf{S}_W \mathbf{w} \mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \mathbf{S}_B \mathbf{w} - \frac{\mathbf{S}_W \mathbf{w} \mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = 0. \quad (5)$$

Let us denote

$$\lambda \triangleq \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}},$$

so that (5) becomes, after moving some terms from left to right,

$$\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}. \quad (6)$$

Multiplying by  $\mathbf{S}_W^{-1}$  (this of course assumes  $\mathbf{S}_W$  to be invertible) we get

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}, \quad (7)$$

which means that  $\mathbf{w}$  has to be an eigenvector of  $\mathbf{S}_W^{-1}\mathbf{S}_B$ . Let us look at the product  $\mathbf{S}_B\mathbf{x}$  for an arbitrary  $\mathbf{x}$ ; substituting the definition (2) we get

$$\mathbf{S}_B\mathbf{x} = (\mathbf{m}_{+1} - \mathbf{m}_{-1})(\mathbf{m}_{+1} - \mathbf{m}_{-1})^T\mathbf{x} = \alpha(\mathbf{x})(\mathbf{m}_{+1} - \mathbf{m}_{-1}),$$

where  $\alpha(\mathbf{x}) = (\mathbf{m}_{+1} - \mathbf{m}_{-1})^T\mathbf{x}$ ;  $\alpha$  is a scalar, which means that for any  $\mathbf{x}$ , the vector  $\mathbf{S}_B\mathbf{x}$  is colinear with  $(\mathbf{m}_{+1} - \mathbf{m}_{-1})$ . We can now go back to (7), and rewrite it

$$\mathbf{S}_W^{-1}\alpha(\mathbf{w})(\mathbf{m}_{+1} - \mathbf{m}_{-1}) = \lambda\mathbf{w}. \quad (8)$$

From here, we immediately get

$$\mathbf{w} = \frac{\lambda}{\alpha(\mathbf{w})}\mathbf{S}_W^{-1}(\mathbf{m}_{+1} - \mathbf{m}_{-1}) \propto \mathbf{S}_W^{-1}(\mathbf{m}_{+1} - \mathbf{m}_{-1}).$$