ELE 535
Machine Learning and Pattern Recognition [1]
Handout #4: Probabilistic Models

Peter J. Ramadge

Fall 2016, version 2.0

Chapter 9

A Quick Probability Review

We assume a basic knowledge of probability and random variables. This chapter gives background on this topic. The material is intended for self review.

9.1 Cumulative Distribution and Density Functions

Let X be a real valued random variable. The cumulative distribution function (CDF) of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x). \tag{9.1}$$

It follows from the definition that if $x \leq y$, then $F_X(x) \leq F_X(y)$. Hence $F_X(x)$ is a nondecreasing function of x. In addition, $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. We also note that for $a < b$

$$F_X(b) - F_X(a) = P(a < X \leq b). \tag{9.2}$$

Hence the CDF defines the probability of X taking a value in any interval $(a, b]$ via the formula (9.2).

Affine Scaling

Let X be real random variable with CDF $F_X(x)$. We often want to consider scaled and shifted versions of X. This is done by defining the new random variable $U = \alpha X + \beta$ as an affine function of X. We restrict attention here to the situation where $\alpha > 0$. It is easy to see that

$$P(U \leq u) = P\left(X \leq \frac{u - \beta}{\alpha}\right).$$

Hence the CDF of U is

$$F_U(u) = F_X\left(\frac{u - \beta}{\alpha}\right). \tag{9.3}$$

9.1.1 Density Functions

If $F_X(x)$ is a differentiable function, then we define the density of X to be

$$f_X(x) = \frac{d}{dx} F_X(x). \tag{9.4}$$

The nondecreasing property of $F_X(x)$ implies that $f_X(x) \geq 0$ and the fundamental theorem of calculus gives the additional properties

$$\int_a^b f_X(x)dx = F_X(b) - F_X(a) = P(a < X \leq b),$$
$$\int_{-\infty}^{x} f_X(s)ds = F_X(x),$$
$$\int_{-\infty}^{\infty} f_X(x)dx = 1.$$

The first of the above properties suggests a useful heuristic interpretation of the density function. Namely that $f_X(x)dx$ is the probability that X takes a value in an infinitesimal interval of width dx at x:

$$f_X(x)dx \approx P(X \in (x - dx, x]).$$

Affine Scaling

Let X be random variable with density $f_X(x)$. For $\alpha, \beta \in \mathbb{R}$ with $\alpha > 0$, form a new random variable $U = \alpha X + \beta$.

Lemma 9.1.1. The random variable $U = \alpha X + \beta$, where $\alpha, \beta \in \mathbb{R}$ with $\alpha > 0$, has the density

$$f_U(u) = \frac{1}{\alpha} f_X\left(\frac{u - \beta}{\alpha}\right). \tag{9.5}$$

Proof. By (9.3), $F_U(u) = F_X\left(\frac{u-\beta}{\alpha}\right)$. Taking the derivative of both sides with respect to u yields (9.5). □

9.2 Expected Value, Mean and Variance

Let X be a real valued random variable with density $f_X(x)$. We can use $f_X(x)$ to find the probability-weighted average of any (well behaved) function $h(X)$ of the random variable. This is called the *expected value* of $h(X)$ and is defined by

$$E[h(X)] = \int_{\mathbb{R}} h(x) f_X(x) dx. \tag{9.6}$$

Important examples include the *mean* (first moment), *second moment*, and *variance* of X. These are defined by:

$$\begin{array}{lrl} \text{mean:} & \mu_X = E[X] & = \int_{-\infty}^{\infty} x f_X(x) dx \\ \text{second moment:} & E[X^2] & = \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ \text{variance:} & \sigma_X^2 = E[(X - \mu_x)^2] & = \int_{-\infty}^{\infty} (x - \mu_x)^2 f_X(x) dx. \end{array} \tag{9.7}$$

By expanding $(X - \mu_x)^2$ it is easily seen that $E[(X - \mu_x)^2] = E[X^2] - \mu^2$.

Affine Scaling

Let X be a random variable with density $f_X(x)$, mean μ_X and variance σ_X^2. Consider $Y = \alpha X + \beta$ formed as an affine function of X with $\alpha > 0$. Although we can find the density of Y, this is not needed to determine it's mean and variance. We have

$$\mu_Y = E[\alpha X + \beta] = \alpha\mu_X + \beta \tag{9.8}$$

$$\sigma_Y^2 = E[(\alpha X + \beta - (\alpha\mu_X + \beta))^2] = E[\alpha^2(X - \mu_X)^2] = \alpha^2\sigma_X^2. \tag{9.9}$$

Example 9.2.1 (The Gaussian Density)**.** A real valued random variable X with density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, \tag{9.10}$$

is said to have a Gaussian density. A simple computation verifies that X has zero mean and unit variance.

Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$ and let $Y = \sigma X + \mu$. Then by Lemma 9.1.1, Y is a real valued random variable with density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}. \tag{9.11}$$

Some elementary computation shows that the density (9.11) has mean μ, variance σ^2, and second moment $\mu^2+\sigma^2$. Alternatively, we can obtain these results using (9.8) and (9.9). This is called the Gaussian density with mean μ and variance σ^2.

9.3 Vector Random Variables

We are most often interested in several random variables. For example, a vector of random variables, a sequence of random variables (random signal), or an array of random variables (a random image).

We will begin with the simplest situation in which we have just two random variables X and Y with densities $f_X(x)$ and $f_Y(y)$, respectively. These densities define each variable's individual statistical characteristics, but do not capture potential dependencies between the two variables. For that we need to consider X and Y jointly. To do so, it is often convenient to group X and Y into a random vector $Z = (X, Y)$.

9.3.1 Joint CDF and Density

To fully describe the random vector Z we need to specify the *joint CDF*

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

or the *joint density*

$$f_{X,Y}(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_{X,Y}(x, y).$$

For every $x_1 < x_2$ and $y_1 < y_2$,

$$\int_{x_1}^{x_2}\int_{y_1}^{y_2} f_{X,Y}(x, y)dxdy = F_{X,Y}(x_2, y_2) - F_{X,Y}(x_1, y_1) = P(x_1 < X \leq x_2, y_1 < Y \leq y_2).$$

This suggests the heuristic interpretation that the probability that X and Y jointly take a value in an infinitesimal rectangle of size $dx \times dy$ at (x, y) is

$$P\,(\, X \in (x - dx, x], Y \in (y - dy, y]\,) \approx f_{X,Y}(x, y)dxdy.$$

9.3.2 Marginal Densities

Let $Z = (X, Y)$ have the joint density $f_{X,Y}(x, y)$. The joint density also specifies the univariate density of each component random variable in Z. These are called the *marginal densities* and are obtained by integrating $f_{X,Y}(x, y)$ over one of the variables:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx.$$

The joint density carries information about statistical dependency between the two components. Whereas the marginal densities only describe the individual random behavior of each component separately.

Let $Z = (X, Y)$ and consider directly computing the mean of Z. This requires computing the mean of each component of Z using $f_{X,Y}(x, y)$. For the first component this yields

$$\int_{y\in\mathbb{R}} \int_{x\in\mathbb{R}} x f_{X,Y}(x, y)dxdy = \int_{x\in\mathbb{R}} x \int_{y\in\mathbb{R}} f_{X,Y}(x, y)dydx = \int_{x\in\mathbb{R}} x f_X(x)dx = \mu_x.$$

Similarly the mean of Y with respect to $f_{X,Y}(x, y)$ is μ_y. If we group the means μ_x, μ_y into a 2-vector $\mu = (\mu_x, \mu_y)$, then we have $E[Z] = \mu = (\mu_x, \mu_y)$. So there is nothing new from the joint density as far as the means are concerned. These can be computed directly from the marginal densities. Similarly, the variances σ_x^2, σ_y^2 can also be computed directly from the marginal densities. These are univariate statistics of each component of Z.

9.3.3 Conditional Densities

Let X and Y be random variables with joint density $f_{XY}(x, y)$. The conditional density of Y given $X = x$ is defined by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)},$$

for all x for which $f_X(x) > 0$. The subscript on the conditional density indicates that Y is being conditioned on X and the argument of f_{XY} gives the dummy variable y that ranges over the possible values of Y and indicates that we are conditioning on X having the particular value x. So the conditional density is formed by fixing the value $X = x$ in the joint density and normalizing the resulting function of y by $f_X(x)$ to ensure that the area under this restricted density is 1. The conditional density thus adjusts our uncertainty in the value assumed by Y given that we know that $X = x$.

From the above definition we see that the joint density can be factored as follows

$$f_{XY}(x, y) = f_{X|Y}(x|y)f_Y(y) \quad \text{and} \quad f_{XY}(x, y) = f_{Y|X}(y|x)f_X(x).$$

9.4 Independent Random Variables

We say that X and Y are *independent* random variables if

$$f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

In this case, the joint density is completely specified by the marginal densities. As a result, there is no statistical dependency between X and Y. When X and Y are not independent, we say that X and Y are *dependent* random variables.

Lemma 9.4.1. If X and Y are independent random variables, then $E[XY] = E[X]E[Y]$.

Proof.

$$\begin{aligned} E[XY] &= \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} xy f_{X,Y}(x,y)dxdy \\ &= \int_{x\in\mathbb{R}} \int_{y\in\mathbb{R}} x f_X(x) y f_Y(y)dxdy \\ &= \int_{x\in\mathbb{R}} x f_X(x)dx \int_{y\in\mathbb{R}} y f_Y(y)dy \\ &= E[X]E[Y]. \end{aligned}$$

□

The converse of Lemma 9.4.1 is not true. It is possible that $E[XY] = E[X]E[Y]$ even though X and Y are dependent random variables.

9.5 Covariance

Determining the average deviation of Z about its mean μ requires considering both direction and magnitude. We can simplify this by first selecting a unit norm (direction) vector $d \in \mathbb{R}^2$ and examining the variance of the real valued random $W = d^T Z = d_1 X + d_2 Y$. This is equivalent to orthogonally projecting Z onto the line in the direction d and considering the variance of the projected random variable.

The mean of W is

$$E[W] = d_1\mu_X + d_2\mu_Y = \mu^T d.$$

To find it's variance we evaluate $E[(W - \mu^T d)^2]$. We can expand $(W - \mu^T d)^2$ as

$$((Z-\mu)^T d)^2 = d^T (Z-\mu)(Z-\mu)^T d.$$

Using this in the expression for the variance of W yields

$$E[(W - \mu^T d)^2] = d^T E[(Z-\mu)(Z-\mu)^T]d.$$

Let's consider the term $E\left[(Z-\mu)(Z-\mu)^T\right]$. Writing the outer product $(Z-\mu)(Z-\mu)^T$ in terms of its components yields the 2×2 matrix

$$\begin{bmatrix} (X-\mu_X)^2 & (X-\mu_X)(Y-\mu_Y) \\ (X-\mu_X)(Y-\mu_Y) & (Y-\mu_Y)^2 \end{bmatrix}$$

and taking the expectation of each term in this matrix with respect to $f_{X,Y}(x,y)$ yields

$$\Sigma_Z = E\left[(Z-\mu)(Z-\mu)^T\right] = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}.$$

The scalar $\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)]$ is called the *covariance* of X and Y and the matrix Σ_Z is called the *covariance matrix* of $Z = (X, Y)$.

Heuristically, the sign and magnitude of σ_{XY} indicate to what extent the deviations of X and Y about their respective means are in phase, out of phase, or incoherent. In the first case, on average $X - \mu_X$ and $Y - \mu_Y$ are both positive or both negative. Hence $E[(X-\mu_X)(Y-\mu_Y)] > 0$. In the second case, on average $X-\mu_X$ and $Y-\mu_Y$ have opposite signs. Hence $E[(X-\mu_X)(Y-\mu_Y)] < 0$. Finally, if $\sigma_{XY} = 0$, then on average there is no coherent behavior between $X - \mu_X$ and $Y - \mu_Y$.

The following lemmas summarize some simple properties of the covariance and the covariance matrix.

Lemma 9.5.1. The covariance of X and Y satisfies $|\sigma_{XY}| \leq \sigma_X \sigma_Y$.

Proof. We can bound σ_{XY} using the Cauchy-Schwartz inequality. From the definition of σ_{XY}:

$$\begin{aligned}|\sigma_{XY}| &= \left| \int_{\mathbb{R}^2} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y) dxdy \right| \\ &\leq \left(\int_{\mathbb{R}^2} (x-\mu_X)^2 f_{X,Y}(x,y) dxdy \right)^{1/2} \left(\int_{\mathbb{R}^2} (y-\mu_Y)^2 f_{X,Y}(x,y) dxdy \right)^{1/2} \\ &= \sigma_X \sigma_Y.\end{aligned}$$

□

Lemma 9.5.2. Σ_Z is symmetric and positive semidefinite. It is positive definite $\Leftrightarrow |\sigma_{XY}| < \sigma_X \sigma_Y$.

Proof. From the equation $\Sigma_Z = E\left[(Z-\mu)(Z-\mu)^T\right]$ we see that $\Sigma_Z^T = \Sigma_Z$. The variance of $W = d^T Z$ is nonnegative and is given by the quadratic function $d^T \Sigma_Z d$. Hence for all unit norm vectors d, $d^T \Sigma_Z d \geq 0$. Thus Σ_Z is positive semidefinite. The proof of the second part relies on a result known as Sylvester's condition. In the case of 2×2 matrices this says that $S = [S_{jk}]$ is positive definite $\Leftrightarrow S_{11} > 0$ and $\det(S) > 0$. In the case at hand, this is equivalent to $\sigma_X^2 > 0$ and $\sigma_X^2 \sigma_Y^2 - \sigma_{XY}^2 > 0$. These conditions are equivalent to $|\sigma_{XY}| < \sigma_X \sigma_Y$. □

9.5.1 Correlated Random Variables

Here we examine the situation $\sigma_{XY} = 0$ more closely.

Lemma 9.5.3. Let X and Y be real valued random variables. Then

(a) X and Y independent $\Rightarrow E[XY] = E[X]E[Y]$.

(b) $E[XY] = E[X]E[Y] \Leftrightarrow \sigma_{XY} = 0$.

(c) X and Y independent $\Rightarrow \sigma_{XY} = 0$.

Proof. Part (a) is the result of Lemma 9.4.1. For part (b) we note that

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] = E[XY] - E[X]E[Y].$$

Hence $\sigma_{XY} = 0$ if and only if $E[XY] = E[X]E[Y]$. For part (c), part (a) implies $E[XY] = E[X]E[Y]$ and then part (b) implies $\sigma_{XY} = 0$. □

The converses of parts (a) and (c) of Lemma 9.5.3 are not true. Dependent random variables X and Y can still have $\sigma_{XY} = 0$. This is because the covariance σ_{XY} is only a second order measure of dependence; it doesn't capture higher order forms of dependence. However, by Lemma 9.5.3, if $\sigma_{XY} \neq 0$, then the random variables must be dependent.

When $\sigma_{XY} = 0$ or equivalently $E[XY] = E[X]E[Y]$ we say that the random variables X and Y are *uncorrelated*. Conversely, when $\sigma_{XY} \neq 0$, we say that X and Y are *correlated*. Note that X and Y being uncorrelated is weaker than X and Y being independent.

Correlation Coefficient

Assume $\sigma_X^2, \sigma_Y^2 > 0$. Then the scalar

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{9.12}$$

is a normalized (dimensionless) measure of the covariance of X and Y. ρ_{XY} is called the *correlation coefficient* or the "Pearson correlation" of X and Y. By Lemma 9.5.1 we have $-1 \leq \rho_{XY} \leq 1$.

9.6 The Bivariate Gaussian Density

We now develop an extension of the Gaussian density to two (and eventually more) random variables. We begin with the easy situation of independent random variables and develop an expression for the density that makes the generalization to dependent random variables obvious.

9.6.1 Independent Components

The simplest instance of a bivariate Gaussian density arises by letting X and Y be independent, zero mean, unit variance, Gaussian random variables. In this case, X and Y have the densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \text{and} \qquad f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \tag{9.13}$$

and the joint density is the product density

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}. \tag{9.14}$$

By setting $Z = (X, Y)$ and $z = (x, y)$ we can rewrite 9.14 as

$$f_{X,Y}(x,y) = f_Z(z) = \frac{1}{2\pi} e^{-\frac{1}{2}z^T z}. \tag{9.15}$$

This simple case of a joint Gaussian density is our starting point.

We can make this more interesting by scaling and shifting X and Y to form new random variables $U = \sigma_U X + \mu_U$ and $V = \sigma_V Y + \mu_V$. Here we assume $\sigma_U, \sigma_V > 0$. Since X and Y are independent random variables, so are U and V.

From Lemma 9.1.1 it follows that U and V have the Gaussian densities

$$f_U(u) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_U} e^{-\frac{1}{2}(u-\mu_U)^2/\sigma_U^2} \qquad \text{and} \qquad f_V(v) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_V} e^{-\frac{1}{2}(v-\mu_V)^2/\sigma_V^2}.$$

So U has mean μ_U and variance σ_U^2 and V has mean μ_V and variance σ_V^2. The independence of U and V then implies that the joint density is the product of the marginal densities:

$$\begin{aligned} f_{U,V}(u,v) &= f_U(u)\, f_V(v) \\ &= \left(\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_U} e^{-\frac{1}{2}(u-\mu_U)^2/\sigma_U^2} \right) \left(\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_V} e^{-\frac{1}{2}(v-\mu_V)^2/\sigma_V^2} \right) \\ &= \frac{1}{2\pi} \frac{1}{\sigma_U \sigma_V} e^{-\frac{1}{2}\left((u-\mu_U)^2/\sigma_U^2 + (v-\mu_V)^2/\sigma_V^2\right)}. \end{aligned}$$

This expression can be simplified if we cast it into vector and matrix notation. Let $W = (U, V)$ be the vector random variable with components U and V. Then let $w = (u, v)$, $\mu = (\mu_U, \mu_V)^T$, and

$$\Sigma = \begin{bmatrix} \sigma_U^2 & 0 \\ 0 & \sigma_V^2 \end{bmatrix}.$$

Notice that Σ is the covariance matrix of X and Y. The argument of the exponential in $f_{U,V}(u,v)$ can now be written as

$$\frac{(u-\mu_U)^2}{\sigma_U^2} + \frac{(v-\mu_V)^2}{\sigma_V^2} = \begin{bmatrix} u-\mu_U & v-\mu_V \end{bmatrix} \begin{bmatrix} \sigma_U^2 & 0 \\ 0 & \sigma_V^2 \end{bmatrix}^{-1} \begin{bmatrix} u-\mu_U \\ v-\mu_V \end{bmatrix} = (w-\mu)^T \Sigma^{-1} (w-\mu).$$

Using this expression and the easily checked equality $|\Sigma|^{1/2} = \sigma_U \sigma_V$, the density can be written in vector form as

$$f_{UV}(u,v) = f_W(w) = \frac{1}{2\pi}\frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)}. \tag{9.16}$$

This density has two parameters: the mean vector $\mu = (\mu_U, \mu_V)$ and the covariance matrix Σ. In the special case consider here, Σ is diagonal and hence is specified by just two parameters, the variances of U and V respectively.

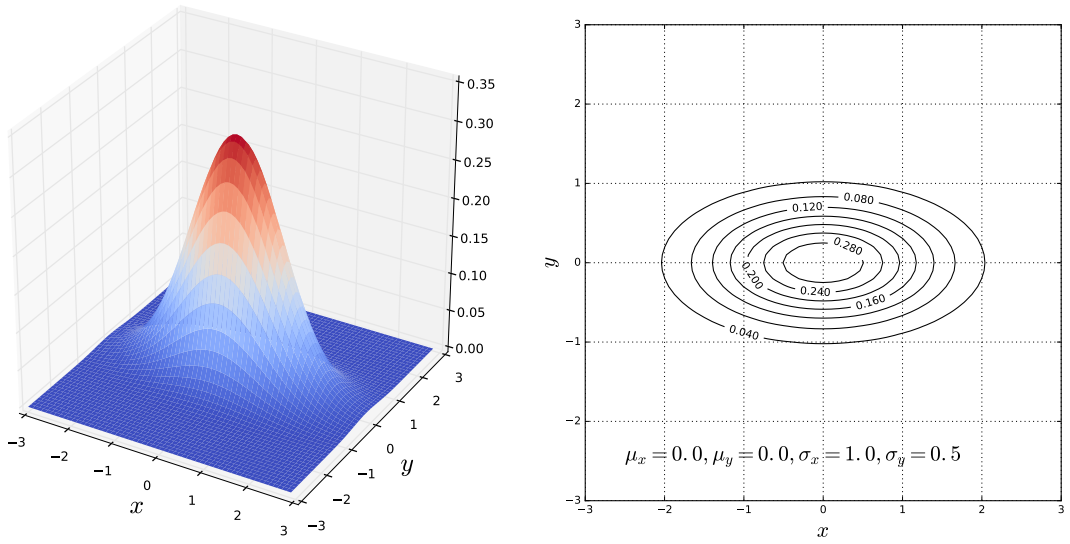

Figure 9.1: A bivariate Gaussian density with independent components. Right: $f_{X,Y}(x,y)$; Left: a contour plot of $f_{X,Y}(x,y)$. When the components are independent and the variances are equal, the contours are circles. When the variances are not equal the contours are ellipses with the major and minor axes aligned with the coordinate axes.

9.6.2 General Form

At this point we suspect the general bivariate Gaussian distribution is given by 9.16 with Σ replaced by a general 2×2 covariance matrix. However, it is instructive to get to this general form in two steps. First consider X and Y to have zero means, unit variances and correlation coefficient $\rho \in (-1, 1)$. So the covariance matrix is

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{9.17}$$

This brings in dependence (when $\rho \neq 0$) in the simplest possible setting. By analogy with (9.16), we let $Z = (X, Y)$, $z = (x, y)$, and define the joint density to be

$$f_{XY}(x,y) = f_Z(z) = \frac{1}{2\pi}\frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} z^T \Sigma^{-1} z\right).$$

Using 9.17 this can be simplified to

$$f_{XY}(x,y) = \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2}\frac{x^2 - 2\rho xy + y^2}{1-\rho^2}\right). \tag{9.18}$$

The marginal densities of X and Y are easily obtained from (9.18) by computing

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}.$$

So X and Y are both Gaussian with (as expected) zero mean and unit variance. Using the fact that X and Y have zero means, the covariance of X and Y can be found by computing the integral

$$E[XY] = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \; e^{-\frac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2)} \, dxdy = \rho$$

So as expected, the covariance of X and Y is ρ. We observe from these computations that when X and Y have a joint Gaussian density the marginal densities are also Gaussian and that when $\rho = 0$, X and Y are independent.

We now scale and shift X and Y to form new random variables $U = \sigma_U X + \mu_U$ and $V = \sigma_V Y + \mu_V$, where $\sigma_U, \sigma_V > 0$. This invertible transformation can be written in matrix form as

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \sigma_U & 0 \\ 0 & \sigma_V \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \sigma_U^{-1} & 0 \\ 0 & \sigma_V^{-1} \end{bmatrix} \left(\begin{bmatrix} U \\ V \end{bmatrix} - \begin{bmatrix} \mu_U \\ \mu_V \end{bmatrix} \right).$$

If we set $S = \begin{bmatrix} \sigma_U & 0 \\ 0 & \sigma_V \end{bmatrix}$. Then the equations relating (x, y) to (u, v) are

$$\begin{bmatrix} x \\ y \end{bmatrix} = S^{-1} \begin{bmatrix} u - \mu_U \\ v - \mu_V \end{bmatrix}.$$

Since $\sigma_U, \sigma_V > 0$, we can then write the joint CDF of (U, V) in terms of the joint CDF of (X, Y) as

$$F_{UV}(u,v) = F_{XY}\left(S^{-1} \begin{bmatrix} u - \mu_U \\ v - \mu_V \end{bmatrix} \right).$$

By taking derivatives with respect to u and v we then obtain

$$f_{U,V}(u,v) = \frac{1}{\sigma_U \sigma_V} f_{X,Y}\left(S^{-1} \begin{bmatrix} u - \mu_U \\ v - \mu_V \end{bmatrix} \right).$$

Writing this out in full yields

$$f_{UV}(u,v) = \frac{1}{2\pi} \frac{1}{|\Sigma|^{1/2}} \frac{1}{\sigma_U \sigma_V} \exp\left(-\frac{1}{2} \begin{bmatrix} u - \mu_U & v - \mu_V \end{bmatrix} S^{-1} \Sigma^{-1} S^{-1} \begin{bmatrix} u - \mu_U \\ v - \mu_V \end{bmatrix} \right).$$

To tidy things up, let $W = (U, V)$ be the vector random variable with components U and V, set $w = (u, v)$, $\mu = (\mu_U, \mu_V)$ and let Ω denote the product $S\Sigma S$. Some simple calculation yields

$$\Omega = \begin{bmatrix} \sigma_U^2 & \sigma_U \sigma_V \rho \\ \sigma_U \sigma_V \rho & \sigma_V^2 \end{bmatrix}$$

and $|\Omega|^{1/2} = \sigma_U \sigma_V (1 - \rho^2) = |\Sigma|^{1/2} \sigma_U \sigma_V$. Thus we can write $f_{UV}(u, v)$ in vector form as

$$f_{UV}(u,v) = f_W(w) = \frac{1}{2\pi} \frac{1}{|\Omega|^{1/2}} e^{-\frac{1}{2}(w-\mu)^T \Omega^{-1} (w-\mu)}. \tag{9.19}$$

Equation 9.19 is the general expression for a bivariate Gaussian density with mean μ and covariance matrix Ω. All together the density has five scalar parameters: the means μ_U, μ_V, the two variances σ_U^2, σ_V^2 and the correlation coefficient $\rho_{UV} = \rho$.

The following lemma summarizes some of the observations made in the above derivations.

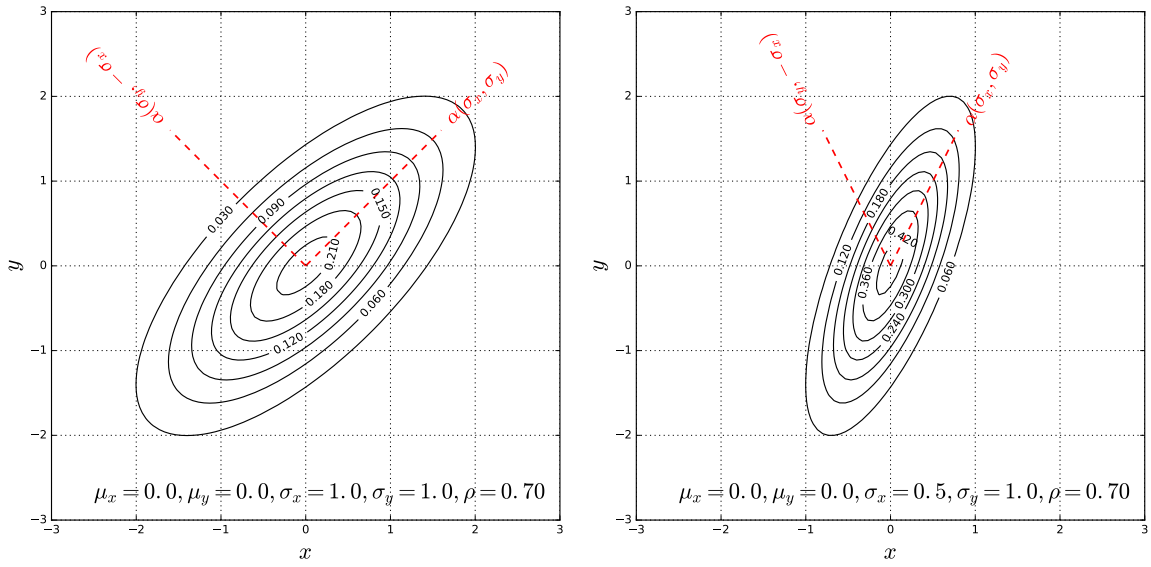

Figure 9.2: Contour plots for bivariate Gaussian densities with dependent components. Right: Equal variances and $\rho = 0.7$. The major axis of the contours is aligned with the line $x = y$ and the contour narrow as $|\rho|$ increases towards 1. Left: distinct variances and $\rho = 0.7$. The major axis of the contours is aligned with the coordinate axis of greatest variance at $\rho = 0$ and then narrows and moves to align with one of the red dashed lines (depending on the sign of ρ) as $|\rho|$ increases towards 1.

Lemma 9.6.1. Let X and Y have a bivariate Gaussian density with mean $\mu = (\mu_X, \mu_Y)^T$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix},$$

with $\sigma_X^2, \sigma_Y^2 > 0$ and $\rho_{XY} \in (-1, 1)$. Then

(a) The marginal densities of X and Y are Gaussian with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively.

(b) X and Y are independent $\Leftrightarrow \rho_{XY} = 0$.

Proof. Exercise. □

Part (b) of the above lemma indicates that if $\rho_{XY} = 0$ (or equivalently $\sigma_{XY} = 0$), then the density (9.19) separates into the product of two one dimensional Gaussian densities and hence X and Y are independent. This is a special property of the bivariate Gaussian that does not hold for general densities.

Chapter 10

The Multivariate Gaussian

10.1 Introduction

We now consider a random vector Z taking values in $\mathbb{R}^n$ for some $n > 1$. We are particularly interested in the situation when Z consists of two dependent component random vectors X and Y. As a motivating application, consider the situation when X generates a data example and Y generates the corresponding label. For this to be interesting, X and Y must be dependent. This dependency is captured in the joint density $f_Z(z) = f_{XY}(x, y)$.

Let X take values in $\mathbb{R}^n$, have density $f_X(x)$, mean μ_X, and covariance Σ_X. Similarly, let Y take values in $\mathbb{R}^q$, have density $f_Y(y)$, mean μ_Y, and covariance Σ_Y. Σ_X describes the second order dependancies among the components of $X - \mu_X$ and Σ_Y does the same for $Y - \mu_Y$. But these do not give us any hint of possible dependency between X and Y. The joint density encodes this dependency. That said, a weaker but nevertheless potentially useful way to model aspects of the dependency is through covariance between the elements X and the elements of Y. This is a component of the covariance of $Z = (X, Y)$. The mean of Z is clearly

$$\mu_Z = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix},$$

and the covariance of Z is

$$E[(Z - \mu_Z)(Z - \mu_Z)^T] = \begin{bmatrix} E[(X - \mu_X)(X - \mu_X)^T] & E[(X - \mu_X)(Y - \mu_Y)^T] \\ E[(Y - \mu_Y)(X - \mu_X)^T] & E[(Y - \mu_Y)(Y - \mu_Y)^T] \end{bmatrix} = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}$$

This brings in two new matrices $\Sigma_{XY} \in \mathbb{R}^{n \times q}$ and $\Sigma_{YX} \in \mathbb{R}^{q \times n}$. Σ_{XY} is called the *cross covariance* of X and Y and Σ_{YX} is called the *cross covariance* of Y and X. Clearly $\Sigma_{XY}^T = \Sigma_{YX}$. Since X and Y need not have the same dimensions, in general Σ_{XY} is not a square matrix.

The conditional density of Y given $X = x$ is defined by

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)},$$

for all x for which $f_X(x) > 0$. Recall that the subscript on the conditional density indicates that Y is being conditioned on X and the argument of f_{XY} gives the dummy variable y that ranges over the possible values of Y and indicates that we are conditioning on X having the particular value x. So the conditional density is formed by fixing the value $X = x$ in the joint density and normalizing the resulting function of y by $f_X(x)$ to ensure that the area under this restricted density is 1. From the definition we see that the joint density can be factored as follows

$$f_{XY}(x, y) = f_{X|Y}(x|y) f_Y(y) \quad \text{and} \quad f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x).$$

10.2 The Multivariate Gaussian Density

Let the random vector X take values in $\mathbb{R}^n$ and have mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$.

We say that X is a *non degenerate Gaussian random vector* if Σ is positive definite and X has the density

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}. \tag{10.1}$$

This is called a *multivariate Gaussian density* and $K = \Sigma^{-1}$ is called the *precision matrix* of the density.

We can use (10.1) to write,

$$\begin{aligned} \ln f_X(x) &= -1/2(x-\mu)^T\Sigma^{-1}(x-\mu) + C \\ &= -1/2\, x^T K x + x^T K\mu + C', \end{aligned} \tag{10.2}$$

where C and C' are constants that do not depend on x. In expression (10.2), the quadratic term specifies K, and the linear term specifies $K\mu$. Since K is known from the quadratic term, this specifies the mean μ. So if $f_X(x)$ is known to be a Gaussian density, then the precision matrix K and the mean μ can be extracted from the quadratic and linear terms in an expansion of $\ln f_X(x)$. Conversely, we have the following lemma.

Lemma 10.2.1. If $f_X(x)$ is a density and $\ln f_X(x)$ has the form (10.2), then $f_X(x)$ is a Gaussian density.

Proof. Exercise. □

10.3 Jointly Gaussian Random Vectors

Consider a Gaussian random vector Z that is partitioned into two component vectors X and Y and concordantly partition the mean and covariance of Z:

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}.$$

We denote the density of Z as $f_{XY}(x,y)$ and the marginal densities of X and Y by $f_X(x)$ and $f_Y(y)$, respectively.

10.3.1 The Conditional Density $f_{X|Y}(x|y)$

Recall that the conditional density of Y given $X = x$ is $f_{Y|X}(y|x) = f_{XY}(x,y)/f_X(x)$. This is a renormalized version of the joint density regarded as a function of y with x fixed.

Theorem 10.3.1. When $f_{XY}(x,y)$ is a Gaussian density, the conditional density $f_{Y|X}(y|x)$ is a Gaussian density with mean $\mu_{Y|X}$ and covariance $\Sigma_{Y|X}$ given by

$$\mu_{Y|X} = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) \tag{10.3}$$

$$\Sigma_{Y|X} = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}. \tag{10.4}$$

Proof. Note that $\ln f_{XY}(x,y) = \ln f_{Y|X}(y|x) + \ln f_X(x)$. Expanding the first term yields

$$\ln f_{XY}(x,y) = -1/2 \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix}^T \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix} + C, \tag{10.5}$$

where C does not depend on x or y. By Lemma 10.4.2 in the Appendix, under the assumption of the existence of the inverses below, we have

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} A^{-1} & \mathbf{0} \\ \mathbf{0} & S_A^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix}.$$

We can use this expression to write (10.5) as

$$-\tfrac{1}{2} \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix}^T \begin{bmatrix} I & -\Sigma_X^{-1}\Sigma_{XY} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \Sigma_X^{-1} & \mathbf{0} \\ \mathbf{0} & S_{\Sigma_X}^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\Sigma_{YX}\Sigma_X^{-1} & I \end{bmatrix} \begin{bmatrix} x-\mu_X \\ y-\mu_Y \end{bmatrix} + C$$

Evaluating the right and left most products first and using the symmetry of Σ yields,

$$\begin{aligned} &-\tfrac{1}{2} \begin{bmatrix} x-\mu_X \\ (y-\mu_Y) - \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) \end{bmatrix}^T \begin{bmatrix} \Sigma_X^{-1} & \mathbf{0} \\ \mathbf{0} & S_{\Sigma_X}^{-1} \end{bmatrix} \begin{bmatrix} x-\mu_X \\ (y-\mu_Y) - \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X) \end{bmatrix} + C \\ &= -1/2(x-\mu_X)^T\Sigma_X^{-1}(x-\mu_X) \\ &\quad -\tfrac{1}{2}(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X))^T S_{\Sigma_X}^{-1}(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)) + C, \end{aligned} \tag{10.6}$$

where $S_{\Sigma_X} = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}$ is the Schur complement of Σ_Y in Σ.

Now fix x and use $\ln f_{XY}(x,y) = \ln f_{Y|X}(y|x) + \ln f_X(x)$ and the above expression to write

$$\ln f_{X|Y}(x|y) = -1/2(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X))^T S_{\Sigma_X}^{-1}(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)) + C + C'$$

where C' do not depend on y but does depend on x. As a function of y, $f_{Y|X}(y|x)$ is a density and $\ln f_{Y|X}(x|y)$ has the form (10.2). Hence by Lemma 10.2.1, $f_{Y|X}(y|x)$ is a Gaussian density. Moreover, (10.3) and (10.4) are obtained directly from the previous expression. □

Notice that the conditional mean $\mu_{Y|X}$ depends on the value x (as expected), but the conditional covariance does not depend on x.

10.3.2 The Marginal Density $f_Y(y)$

The marginal densities $f_X(x)$ and $f_Y(y)$ are obtained by integrating $f_{XY}(x,y)$ over the values of X and Y, respectively.

Lemma 10.3.1. If the joint $f_{XY}(x,y)$ is Gaussian so are the marginal densities $f_X(x)$ and $f_Y(y)$. Moreover, the means and covariances are μ_X, Σ_X and μ_Y, Σ_Y, respectively.

Proof. Using the expression (10.6) for $\ln f_{XY}(x,y)$ yields

$$f_{XY}(x,y) = C\, e^{-1/2(x-\mu_X)^T\Sigma_X^{-1}(x-\mu_X)}\, e^{-1/2(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X))^T S_{\Sigma_X}^{-1}(y-\mu_Y-\Sigma_{YX}\Sigma_X^{-1}(x-\mu_X))}.$$

Integrating this expression over y gives $f_X(x) = CC'e^{-1/2(x-\mu_X)^T\Sigma_X^{-1}(x-\mu_X)}$, where C' does not depend on x or y. This proves the result for $f_X(x)$. The proof for $f_Y(y)$ follows a symmetric argument. □

10.4 Appendix: Matrix Inversion Using the Schur Complement

Let $A \in \mathbb{R}^{p\times p}$ and $D \in \mathbb{R}^{q\times q}$ be square matrices and consider block matrices of the form

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

Assuming A is invertible, we can zero the block below A by left matrix multiplication:

$$\begin{bmatrix} I_p & \mathbf{0} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ \mathbf{0} & D - CA^{-1}B \end{bmatrix}.$$

Similarly, we can zero the block to the right of A by right matrix multiplication:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ \mathbf{0} & I_q \end{bmatrix} = \begin{bmatrix} A & \mathbf{0} \\ C & D - CA^{-1}B \end{bmatrix}.$$

The same matrices operating together yield a block diagonal result:

$$\begin{bmatrix} I_p & \mathbf{0} \\ -CA^{-1} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & -A^{-1}B \\ \mathbf{0} & I_q \end{bmatrix} = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & S_A \end{bmatrix}, \tag{10.7}$$

where $S_A = D - CA^{-1}B$ is called the *Schur complement of A in M*.

The matrices used above are invertible and the inverses are easily found.

Lemma 10.4.1. For all $X \in \mathbb{R}^{q\times p}$ and $Y \in \mathbb{R}^{p\times q}$:

$$\begin{bmatrix} I_p & \mathbf{0} \\ X & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & \mathbf{0} \\ -X & I_q \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I_p & Y \\ \mathbf{0} & I_q \end{bmatrix}^{-1} = \begin{bmatrix} I_p & -Y \\ \mathbf{0} & I_q \end{bmatrix}$$

Proof. Multiply. □

Combining Lemma 10.4.1 with (10.7) yields a simple expression for the inverse of the matrix M.

Lemma 10.4.2. If M and A are invertible, then S_A is invertible and

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} I_p & -A^{-1}B \\ \mathbf{0} & I_q \end{bmatrix} \begin{bmatrix} A^{-1} & \mathbf{0} \\ \mathbf{0} & S_A^{-1} \end{bmatrix} \begin{bmatrix} I_p & \mathbf{0} \\ -CA^{-1} & I_q \end{bmatrix} \\ &= \begin{bmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{bmatrix}. \end{aligned}$$

Proof. If M and A are invertible, then (10.7) and Lemma 10.4.1 imply that S_A^{-1} is invertible. The result then follows by taking the inverse of both sides of (10.7) and using Lemma 10.4.1. □

A parallel set of results can be obtained by assuming that D is invertible. In this case,

$$\begin{bmatrix} I_p & -BD^{-1} \\ \mathbf{0} & I_q \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I_p & \mathbf{0} \\ -D^{-1}C & I_q \end{bmatrix} = \begin{bmatrix} S_D & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix},$$

where $S_D = A - BD^{-1}C$ is the *Schur complement of D in M*. This yields the following result.

Lemma 10.4.3. If M and D are invertible, then S_D is invertible and

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} I_p & \mathbf{0} \\ -D^{-1}C & I_q \end{bmatrix} \begin{bmatrix} S_D^{-1} & \mathbf{0} \\ \mathbf{0} & D^{-1} \end{bmatrix} \begin{bmatrix} I_p & -BD^{-1} \\ \mathbf{0} & I_q \end{bmatrix} \\ &= \begin{bmatrix} S_D^{-1} & -S_D^{-1}BD^{-1} \\ -D^{-1}CS_D^{-1} & D^{-1} + D^{-1}CS_D^{-1}BD^{-1} \end{bmatrix}. \end{aligned}$$

Proof. Exercise. □

Chapter 11

Maximum Likelihood Estimation

Let X be a non degenerate Gaussian random vector with density

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \tag{11.1}$$

Here Σ is symmetric positive definite. If μ and Σ are unknown, then we can try to learn these parameters using a set of training data $\{x_i \in \mathbb{R}^n\}_{i=1}^m$ where each example is drawn independently under $f_X(x)$.

Let

$$L(x_1, \ldots, x_m; \mu, \Sigma) = \prod_{i=1}^m f_X(x_i) = \prod_{i=1}^m \frac{1}{(2\pi)^{\frac{n}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)}.$$

With $x_1, \ldots, x_m$ fixed and μ, Σ the variables, this function is called the *likelihood function*. It measures the likelihood of the observed training data under each set of parameters. In *maximum likelihood estimation* one sets out to estimate the unknown parameters by maximizing the likelihood function, or equivalently by maximizing the log-likelihood $\ln(L)$. In this case, the log-likelihood is

$$\ln(L) = -\frac{1}{2} m \ln \det(\Sigma) - \frac{1}{2} \sum_{i=1}^m (x_i - \mu) \Sigma^{-1} (x_i - \mu) + C,$$

where the constant C does not depend on the data, μ, or Σ. Thus the problem of maximizing the log-likelihood is equivalent to:

$$\begin{aligned} \min_{\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}} \quad & J(\mu, \Sigma) = m \ln \det(\Sigma) + \sum_{i=1}^m (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \\ \text{s.t.} \quad & \Sigma \text{ is symmetric PD.} \end{aligned} \tag{11.2}$$

First set the derivative of $J(\mu, \Sigma)$ with respect to μ equal to zero. This gives

$$\begin{aligned} D_\mu J(\mu, \Sigma)(h) &= \sum_{i=1}^m -h^T \Sigma^{-1}(x_i - \mu) - (x_i - \mu)^T \Sigma^{-1} h \\ &= -2 \left(\sum_{i=1}^m (x_i - \mu)^T \Sigma^{-1} \right) h \\ &= 0. \end{aligned}$$

Since this holds for all $h \in \mathbb{R}^n$, it follows that $\sum_{i=1}^m (x_i - \mu)^T \Sigma^{-1} = \mathbf{0}$. Multiplying both sides of this expression by Σ and rearranging gives the maximum likelihood estimate

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^m x_i. \tag{11.3}$$

This is just the empirical mean of the training data. Note that this expression does not depend of Σ.

We can now substitute $\hat{\mu}$ for μ in $J(\mu, \Sigma)$ to obtain a new objective that is only a function of Σ. It is convenient to do this by setting $z_i = x_i - \hat{\mu}$ and $S = \sum_{i=1}^m z_i z_i^T$. Noting that $z_i^T \Sigma^{-1} z_i = \mathrm{trace}(z_i z_i^T \Sigma^{-1})$, gives $\sum_{i=1}^m z_i^T \Sigma z_i = \mathrm{trace}(S\Sigma^{-1})$. The new problem can now be written as

$$\begin{aligned} \min_{\Sigma \in \mathbb{R}^{n\times n}} \quad & J(\Sigma) = m \ln \det(\Sigma) + \mathrm{trace}(S\Sigma^{-1}) \\ \text{s.t.} \quad & \Sigma \text{ is symmetric PD.} \end{aligned} \tag{11.4}$$

The symmetric matrices in $\mathbb{R}^{n\times n}$ form a subspace of dimension $(n+1)n/2$. The symmetric positive semidefinite matrices constitute a closed subset C of this subspace and the symmetric positive definite matrices form the interior of C. If 11.2 has a positive definite solution, then this lies in the interior of C and we can use calculus to try to find it.

To take the derivative of the objective function in (11.4) with respect to Σ we first determine the derivatives of the functions $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ with $f(M) = \det(M)$ and $g\colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ with $g(M) = M^{-1}$ (where M is assumed to be invertible). Expressions for these derivatives for general M (not necessarily symmetric) are given in the following two lemmas.

Lemma 11.0.4. For any invertible $M \in \mathbb{R}^{n\times n}$, let $g(M) = M^{-1}$. Then

$$Dg(M)(H) = -M^{-1} H M^{-1}. \tag{11.5}$$

Proof. We use the equality $g(M)M = I_n$. Taking the derivatives of both sides of this expression yields $Dg(M)(H)M = -g(M)H$. Hence $Dg(M)(H) = -M^{-1}HM^{-1}$. □

So if we are at M and move a very small step ϵ along H to $M + \epsilon H$, then to first order in ϵ:

$$(M + \epsilon H)^{-1} = M^{-1} + Dg(M)(\epsilon H) = M^{-1} - \epsilon M^{-1} H M^{-1}.$$

Lemma 11.0.5. For $M \in \mathbb{R}^{n\times n}$, let $f(M) = \det(M)$. Then for any invertible M,

$$Df(M)(H) = \det(M)\,\mathrm{trace}(M^{-1}H). \tag{11.6}$$

Proof. From the definition of the derivative

$$\begin{aligned} Df(M)(H) &= \lim_{\epsilon\to 0} \frac{f(M+\epsilon H) - f(M)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{\det(M(I + \epsilon M^{-1}H)) - \det(M)}{\epsilon} \\ &= \det(M) \lim_{\epsilon\to 0} \frac{\det(I + \epsilon M^{-1}H)) - 1}{\epsilon} \end{aligned}$$

To proceed, we need to evaluate $\det(I + \epsilon A)$ to first order in epsilon, where $A = M^{-1}H$. Here we give an intuitive derivation, a more detailed proof is given in the appendix. For ϵ very small, the value of

$\det(I + \epsilon A)$ is dominated by the diagonal elements of $I + \epsilon A$. Ignoring the off diagonal elements and focusing on terms that are first order in ϵ, we have

$$\det(I + \epsilon A) = \prod_{i=1}^{n}(1 + \epsilon A_{ii}) = 1 + \epsilon \operatorname{trace}(A).$$

Thus

$$Df(M)(H) = \det(M) \lim_{\epsilon \to 0} \frac{\det(I + \epsilon M^{-1}H)) - 1}{\epsilon} = \det(M) \operatorname{trace}(M^{-1}H).$$

□

In this case, if we are at M and move a very small step ϵ along H to $M + \epsilon H$, then to first order in ϵ:

$$\det(M + \epsilon H) = \det(M) + Df(M)(\epsilon H) = \det(M) + \epsilon \det(M) \operatorname{trace}(M^{-1}H).$$

Now we return to (11.4) and set the derivative of the objective function with respect to Σ equal to zero. This gives

$$\begin{aligned} DJ(\Sigma)(H) &= m\frac{1}{\det(\Sigma)} \det(\Sigma) \operatorname{trace}(\Sigma^{-1}H) - \operatorname{trace}(S\Sigma^{-1}H\Sigma^{-1}) \\ &= \operatorname{trace}\left(\left(m\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}\right)H\right) \\ &= 0 \end{aligned}$$

Thus for all H, $<m\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}, H> = 0$. It follows that $m\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1} = \mathbf{0}$. Multiplying both sides of this expression on the right and left by Σ and rearranging yields the candidate maximum likelihood estimate

$$\hat{\Sigma} = \frac{1}{m}S = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^T. \tag{11.7}$$

This estimate is just the empirical covariance of the training data. It is symmetric and positive semidefinite but it might fail to be positive definite. Assuming $f_X(x)$ is non degenerate, if $\hat{\Sigma}$ fails to be positive definite, then we have used insufficient training data.

We have proved the following Theorem.

Theorem 11.0.1. Let $\{x_i\}_{i=1}^m$ be independent samples from a non degenerate multivariate Gaussian density with mean $\mu \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$. Then the maximum likelihood estimates of μ and Σ based on the given samples are

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{\mu})(x_i - \hat{\mu})^T.$$

11.1 Learning a Parameter of the Mean

Let $B \in \mathbb{R}^{p \times n}$ be given, $w \in \mathbb{R}^n$ be fixed but unknown, and V be a Gaussian random vector with covariance Σ. Let $Y = Bw + V$. So Y is a Gaussian random vector with mean $\mu = Bw$, covariance Σ, and density

$$f_Y(y; w) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} e^{-1/2(y-Bw)^T\Sigma^{-1}(y-Bw)}.$$

We have written $f_Y(y; w)$ as a function of both w and y. For fixed w, it is the density of Y; for fixed y, it is the likelihood function $L(w)$ of w given the observation y.

In the framework of the above model, we want to use the information provided by the noisy measurement $y = Aw + v$ to obtain the maximum likelihood estimate $\hat{w}$ of w. Equivalently, one can maximize the the log-likelihood formed by taking natural log of the likelihood function. Assuming B and Σ are known, maximizing the log-likelihood requires:

$$\begin{aligned}\hat{w} &\in \arg\max_{w\in\mathbb{R}^n} -1/2(y-Bw)^T\Sigma^{-1}(y-Bw)\\ &= \arg\min_{w\in\mathbb{R}^n} (y-Bw)^T\Sigma^{-1}(y-Bw)\\ &= \arg\min_{w\in\mathbb{R}^n} \|y-Bw\|^2_{\Sigma^{-1}} \qquad (11.8)\\ &= \arg\min_{w\in\mathbb{R}^n} \|y-Bw\|_2^2 \quad \text{if } \Sigma = \sigma^2 I_p \qquad (11.9)\end{aligned}$$

Equation (11.8) is a variation on a least squares problem. The special case (11.9), assumes the measurement noise has i.i.d. components, and yields a standard least squares problem. So for the assumed model, if the noise covariance is known or is known to be i.i.d. , then the maximum likelihood estimate of w given y is obtained by solving a least squares problem.

11.2 Appendix: Additional Proofs

11.2.1 The Determinant of $I + \epsilon A$

Lemma 11.2.1. For $A \in \mathbb{R}^{n\times n}$ and small $\epsilon \in \mathbb{R}$, $|I + \epsilon A| = 1 + \epsilon\,\mathrm{trace}(A) + O(\epsilon^2)$.

Proof. If A is a 2×2 matrix, then

$$|I+\epsilon A| = \begin{vmatrix} 1+\epsilon a_{11} & \epsilon a_{12} \\ \epsilon a_{21} & 1+\epsilon a_{22}\end{vmatrix} = 1 + \epsilon\,\mathrm{trace}(A) + \epsilon^2|A|.$$

So the result holds for 2×2 matrices.

Assume the result holds for any real $n \times n$ matrix. We show it then holds for any real $(n+1)\times(n+1)$ matrix. Let B be an $(n+1)\times(n+1)$ matrix and write B in the form

$$B = \begin{bmatrix} A & b \\ c^T & a_{n+1,n+1}\end{bmatrix}$$

with $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}^n$ and $a_{n+1,n+1} \in \mathbb{R}$. Then

$$|I+\epsilon B| = \begin{vmatrix} I+\epsilon A & \epsilon b \\ \epsilon c^T & 1+\epsilon a_{n+1,n+1}\end{vmatrix}.$$

Expand the determinant along the last row of the matrix. This yields a sum of $n+1$ terms. The terms from the first n entries of the last row have the form $(-1)^{(n+1)i}\epsilon c_i|D_i|$, where D_i is the submatrix formed by eliminating the last row and i-th column of the matrix $I + \epsilon B$, $i = 1, \ldots, n$. Notice that the last column of D_i is ϵb. So $|D_i| = \epsilon|\bar{D}_i|$, where $\bar{D}_i$ is obtained from D_i by replacing ϵb by b. Here we have used the property of the determinant that scaling one column scales the determinant. Hence each of the first n terms has the form $(-1)^{(n+1)i}\epsilon^2 c_i|\bar{D}_i|$. Now $|\bar{D}_i|$ is a sum of products of entries of $\bar{D}_i$ and hence is a polynomial in ϵ. Hence each of the first n terms is $O(\epsilon^2)$. Now let's look at the last term in the expansion. Using the induction hypothesis, this has the form

$$(1+\epsilon a_{n+1,n+1})|I+\epsilon A| = (1+\epsilon a_{n+1,n+1})(1+\epsilon\,\mathrm{trace}(A)+O(\epsilon^2)) = 1+\mathrm{trace}(B)+O(\epsilon^2)$$

Thus $|I + \epsilon B| = I + \epsilon\,\mathrm{trace}(B) + O(\epsilon^2)$. □

11.2.2 The Derivative $(D|A|)H$

Lemma 11.2.2. For invertible $A \in \mathbb{R}^{n\times n}$, and any $H \in \mathbb{R}^{n\times n}$,

$$(D_A|A|)H = |A| \operatorname{trace}(A^{-1}H). \tag{11.10}$$

Proof. Using Lemma 11.2.1

$$\begin{aligned}\frac{|A+\epsilon H| - |A|}{\epsilon} &= \frac{|A(I+\epsilon A^{-1}H)| - |A|}{\epsilon} \\ &= \frac{|A|(1+\epsilon \operatorname{trace}(A^{-1}H) + O(\epsilon^2)) - |A|}{\epsilon} \\ &= |A| \operatorname{trace}(A^{-1}H) + \frac{O(\epsilon^2)}{\epsilon}.\end{aligned}$$

Hence

$$\lim_{\epsilon\to 0} \frac{|A+\epsilon H| - |A|}{\epsilon} = |A| \operatorname{trace}(A^{-1}H).$$

□

11.3 Exercises

Exercise 11.1. Consider the problem in §11.1. Suppose we have m independent samples $\{y_i\}_{i=1}^m$ and $\Sigma = \sigma^2 I_p$. Let $\bar{y} = \frac{1}{m}\sum_{i=1}^m y_i$. Show that maximum likelihood solution is obtained by solving

$$\hat{w} = \arg\min_{w\in\mathbb{R}^n} \|\bar{y} - Bw\|_2^2$$

If Σ is a general positive definite matrix, show that the maximum likelihood solution is obtained by solving

$$\hat{w} = \arg\min_{w\in\mathbb{R}^n} \|\bar{y} - Bw\|_{\Sigma^{-1}}^2$$

Chapter 12

Estimating the Value of a Random Vector

12.1 Introduction

We now consider the problem of estimating (or inferring) the value of a random vector Y given the outcome of a related random vector X. Generally, the inferred value of Y will be a function of the observed value of X. We refer to this function as an *estimator*. In contrast, the value produced by the estimator for a particular observed value x of X is called an *estimate* of Y. A situation of particular interest is when X is a data vector that we can observe (measure), and Y is an unobserved label vector that we want to predict. We consider the general situation where X takes values in $\mathbb{R}^n$ and Y takes values in $\mathbb{R}^q$, and initially assume that we know the joint density of X and Y, or at least the first and second order statistics of this density. The special case when $q = 1$ makes immediate contact with our prior discussion of least squares regression.

We begin by considering how to estimate the value of a random vector Y given its density. Then we consider estimating the value of Y given the joint density $f_{XY}(x, y)$ and the value assumed by X. Finally we re-examine the second scenario but this time restrict the family of allowed estimators to be the set of affine functions of x.

12.2 Estimate the Value of Y Given its Density $f_Y(y)$

Let Y be a random vector with density $f_Y(y)$. Using only this information we want to infer (or estimate) the value assumed by Y. Naturally this will incur some error. So we seek a value $\hat{y}$ (the estimate) that minimizes an appropriate cost function of the error $Y - \hat{y}$. One possibility is to minimize the expected value of the squared error, $E[\|Y - \hat{y}\|_2^2]$. This is called the *mean squared error* (MSE) cost. To ensure that each $\hat{y}$ yields a finite MSE cost, we will assume that Y has a finite first and second moments. Then

$$E[\|Y - \hat{y}\|_2^2] = E[\|Y - \mu_Y + \mu_Y - \hat{y}\|_2^2] \leq E[\|Y - \mu_Y\|_2^2] + \|\mu_Y - \hat{y}\|_2^2 = \text{trace}(\Sigma_Y) + \|\mu_Y - \hat{y}\|_2^2.$$

Minimizing the MSE cost leads to the intuitively reasonable result that the best MSE estimate of Y is its mean μ_Y.

Lemma 12.2.1. Assume that the components of Y have finite first and second order moments. Then $\hat{y} = \mu_Y$ minimizes the MSE $E[(Y - \hat{y})^2]$.

Proof. We want to set the gradient of $E[(Y - \hat{y})^2]$ with respect to $\hat{y}$ equal to zero. To do so, we first exchange the order of the gradient and expectation operators. Our assumptions on Y ensure that this is possible (see the appendix). Setting the expected value of the gradient equal to zero gives $2E[-Y+\hat{y}] = 0$. Hence the optimal selection is $\hat{y} = E[Y] = \mu_Y$. □

Had we chosen a different cost function we would get a different answer for $\hat{y}$ but it will still be some constant determined by the density of Y and the selected cost function. Since the MSE cost function is quadratic in $\hat{y}$, it is simple to work with and it yields an intuitively reasonable result.

12.3 Estimate the Value of Y Given $f_{XY}(x, y)$ and $X = x$

We now consider how to estimate the value of Y when we have additional information available. To model this situation bring in a second random variable X and let X and Y have a joint density $f_{XY}(x, y)$. We suppose that a realization of (X, Y) is determined but we only measure the value assumed by X, i.e., we know that $X = x$. Assuming X and Y are dependent, this information enables a more accurately inference of the value of Y.

Lemma 12.3.1. Assume the conditional density $f_{Y|X}(y|x)$ has finite first and second order moments. Then the optimal MSE estimate of the value of Y given that $X = x$ is the mean of the conditional density $f_{Y|X}(y|x)$:

$$\hat{y}(x) = E[Y|X = x] = \int_{\mathbb{R}} y f_{Y|X}(y|x) dy = \mu_{Y|X}(x).$$

Proof. The conditional density $f_{X|Y}(y|x)$ adjusts the density of Y to fully exploit the information provided by the observation $X = x$. Any residual uncertainty in the value of Y is completely described by $f_{X|Y}(y)$. Under the assumptions of By Lemma 12.2.1, the optimal MSE estimate of the value of Y is the mean of the conditional density. □

The mean of the condition density is a function of the observed value x of X. Hence it is an estimator. The value of the conditional mean for a particular observed value x of X is the corresponding estimate $\hat{y}$ of Y. In general, it is difficult to find an expression for the (generally nonlinear) conditional mean estimator. So although we know that this estimator exists and that it is optimal for the MSE cost, applying this knowledge to compute the estimator is a challenge.

12.3.1 Conditional Mean Estimator for Jointly Gaussian Random Vectors

One case when the conditional mean estimator is easily computable is when $f_{XY}(x, y)$ is Gaussian. In this case the condition mean was previous shown to be

$$\hat{y}(x) = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x - \mu_X).$$

This is an affine function of x: $\hat{y}(x) = Wx + b$ where $W = \Sigma_{YX}\Sigma_X^{-1}$ and $b = (\mu_Y - W\mu_X)$.

For simplicity, consider the situation when $q = 1$ and Y is a scalar random variable. In that case W is a row vector, denoted henceforth by w^T, and the MSE estimator is $\hat{y}(x) = w^T x + b$. The vector w is the solution of $w^T\Sigma_X = \Sigma_{YX}$, or equivalently of $\Sigma_X w = \Sigma_{XY}$.

The above assumes we know Σ_X and Σ_{XY}. Suppose instead we are given training data $\{x_i, y_i\}_{i=1}^m$. Let A denote the matrix with the training examples as its rows and y denote the vector containing the corresponding scalar labels. Then the the maximum likelihood estimate of Σ_X is $\frac{1}{m}A^T A$ and of Σ_{XY} is $\frac{1}{m}A^T y$. If we use the maximum likelihood estimates in place of Σ_X and Σ_{XY} to determine w, then we seek the solution of $A^T Aw = A^T y$. Such a w will be recognized as the solution of the least squares problem $\min_{w\in\mathbb{R}^n} \|y - Aw\|_2^2$. So under the assumption of a Gaussian joint density, solving the least squares problem is equivalent to using the training data to obtain the maximum likelihood estimates of the relevant terms in the covariance matrix and using these estimates in the condition mean estimator.

12.3.2 Using a Fixed Affine Function to Estimate Y Given $X = x$

An alternative to finding the condition mean estimator $\mu_{Y|X}(x)$ is to seek the best estimator over a given family of estimators. Here we consider estimators that are affine functions of x, i.e., $\hat{y} = Wx + b$ where $W \in \mathbb{R}^{q\times n}$ and $b \in \mathbb{R}^q$ do not depend on x. In this approach we want to find a matrix $W \in \mathbb{R}^{q\times n}$ and a vector $b \in \mathbb{R}^q$ such that averaged over the values of X and Y, $\hat{y}(x) = Wx + b$ yields a minimum MSE estimate of Y. This can be stated as

$$\min_{W\in\mathbb{R}^{q\times n},\, b\in\mathbb{R}^q} E\left[\sum_{j=1}^{q}(Y_j - (WX)_j - b_j)^2\right] = E\left[\|Y - WX - b\|_2^2\right]. \tag{12.1}$$

Let X have mean $\mu_X \in \mathbb{R}^n$ and covariance $\Sigma_X \in \mathbb{R}^{n\times n}$, Y have mean $\mu_Y \in \mathbb{R}^p$ and covariance $\Sigma_Y \in \mathbb{R}^{p\times p}$, and let the cross covariance of X and Y be $\Sigma_{XY} \in \mathbb{R}^{n\times p}$. A solution of (12.1) can be found using only these quantities.

Theorem 12.3.1. The minimum MSE affine estimator of Y given $X = x$ is

$$\hat{y}(x) = W^\star(x - \mu_X) + \mu_Y, \tag{12.2}$$

where $W^\star = \Sigma_{YX}\Sigma_X^{-1}$ if Σ_X is invertible and is any solution of $W^\star\Sigma_X = \Sigma_{YX}$ otherwise.

Proof. Assume for the moment that $\mu_X = \mathbf{0}$ and $\mu_Y = \mathbf{0}$. Expanding the RHS of (12.1) yields

$$\begin{aligned} E\left[\|Y - WX - b\|_2^2\right] &= E\left[(Y - WX - b)^T(Y - WX - b)\right] &(12.3)\\ &= E\left[(Y - WX)^T(Y - WX)\right] + E\left[-2(Y - WX)^Tb + b^Tb\right] \\ &= E\left[(Y - WX)^T(Y - WX)\right] + b^Tb. &(12.4)\end{aligned}$$

Since both of the terms in (12.4) are nonnegative and the first does not depend on b, it follows that the expression is minimized with $b = \mathbf{0}$. Using the properties of the trace function, the first term in (12.4) can rewritten as

$$\begin{aligned} E\left[(Y - WX)^T(Y - WX)\right] &= E\left[\mathrm{trace}(Y^TY - Y^TWX - X^TW^TY + X^TW^TWX)\right] \\ &= E\left[\mathrm{trace}\left(YY^T - 2WXY^T + WXX^TW^T\right)\right] &(12.5)\end{aligned}$$

Hence

$$E\left[\|Y - WX - b\|_2^2\right] = \mathrm{trace}\left(\Sigma_Y - 2W\Sigma_{XY} + W\Sigma_XW^T\right). \tag{12.6}$$

The derivative of (12.6) with respect to A acting on $H \in \mathbb{R}^{n\times n}$ is

$$\mathrm{trace}\left(-2H\Sigma_{XY} + H\Sigma_XW^T + W\Sigma_XH^T\right) = 2\,\mathrm{trace}\left(H\Sigma_XW^T - H\Sigma_{XY}\right). \tag{12.7}$$

Setting this expression equal to zero we find that for every $H \in \mathbb{R}^{n\times n}$:

$$\mathrm{trace}\left(H^T(W\Sigma_X - \Sigma_{YX})\right) = <H, W\Sigma_X - \Sigma_{YX}> = 0.$$

It follows that a necessary condition for W to minimize the MSE cost is that

$$W\Sigma_X = \Sigma_{YX}. \tag{12.8}$$

To verify that a solution of (12.8) minimizes the objective function we can compute the second derivative, i.e., the derivative with respect to W of the RHS of (12.7). This yields $\mathrm{trace}(H\Sigma_XH^T)$. Since H is any

$n \times n$ matrix we can replace it by H^T to obtain the equivalent expression $\text{trace}(H^T \Sigma_X H)$. Noting that Σ_X is PSD, we conclude that $\text{trace}(H^T \Sigma_X H) \geq 0$ and hence that all solutions of (12.8) minimize the MSE cost objective. If Σ_X is positive definite, (12.7) has the unique solution

$$W^\star = \Sigma_{YX} \Sigma_X^{-1}. \tag{12.9}$$

When the means μ_X and μ_Y are not zero, we apply the reasoning above to predict $Y - \mu_Y$ given the value of $X - \mu_X$. This yields the minimum MSE predictor $W^\star(x - \mu_X)$ with $W^\star$ given by (12.9). Hence the least MSE predictor of Y is $\hat{y}(x) = W^\star(x - \mu_X) + \mu_Y$, when Σ_X is invertible or any solution of (12.8) otherwise. □

12.4 Application to Label Prediction

Suppose we observe a data vector X taking values in $\mathbb{R}^n$ and want to estimate a corresponding label vector Y taking values in $\mathbb{R}^q$. Set $Z = (X, Y)$ and assume that we know μ_Z and Σ_Z:

$$\mu_Z = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \qquad \Sigma_Z = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix}.$$

Then we can use the minimum MSE affine estimator (12.2) to infer the value of the label vector Y given that $X = x$:

$$\hat{y}(x) = \Sigma_{YX} \Sigma_X^{-1}(x - \mu_X) + \mu_Y = W^\star(x - \mu_X) + \mu_Y$$

where $W^\star = \Sigma_{YX} \Sigma_X^{-1}$ is a $q \times n$ matrix. This can be rewritten as

$$\hat{y}(x) - \mu_Y = W^\star(x - \mu_X).$$

If the joint density of the examples and labels is Gaussian, then this is the optimal MSE estimator.

12.4.1 Special Case: Label Denoising

Suppose that Y is a random label vector with mean μ_Y and covariance Σ_Y. However, instead of directly observing the outcomes of Y, we observe (or measure) $X = Y + N$ where N is a random noise vector of the same size as Y. So X is observed while Y and N are latent (unobserved) random vectors. We assume that the noise vector N has zero mean, covariance Σ_N, and is independent of Y. We want to use the observation $X = x$ and knowledge of the first and second order statistics of both Y and the noise N to infer the value of Y. To do so, we adopt the MSE cost criterion and use the affine estimator (12.2).

It is easily checked that

$$\begin{aligned} \mu_X &= E[Y + N] = \mu_Y \\ \Sigma_X &= E[(Y - \mu_Y + N)(Y - \mu_Y + N)^T] = \Sigma_Y + \Sigma_N \\ \Sigma_{XY} &= E[(Y - \mu_Y + N)(Y - \mu_Y)^T] = \Sigma_Y. \end{aligned}$$

Hence by (12.2) the minimum MSE affine estimate of the value of Y given the observation $X = x$ is

$$\hat{y}(x) = \Sigma_Y(\Sigma_Y + \Sigma_N)^{-1}(x - \mu_Y) + \mu_Y \tag{12.10}$$

Additional insight can be obtained by considering the special case when the noise components are uncorrelated and each has the same variance σ_N^2, i.e., $\Sigma_N = \sigma_N^2 I_q$. Let $\Sigma_Y = UDU^T$ be an eigen-decomposition of the symmetric positive semidefinite matrix Σ_Y. So the columns u_j, $j = 1, \ldots, q$, of U

are orthonormal eigenvectors of Σ_Y and D denotes the diagonal matrix formed from the eigenvalues σ_j^2, $j = 1, \ldots, q$. The random variable $u_j^T(Y - \mu_Y)$ is the projection of $Y - \mu_Y$ onto the line in the direction u_j. The variance of this random variable is

$$E[u_j^T(Y - \mu_Y)(Y - \mu_Y)^T u_j] = u_j^T \Sigma_Y u_j = \sigma_j^2.$$

So we can interpret σ_j^2 as the variance of Y in the direction of its j-th unit norm eigenvector u_j.

Now use the fact that $U^T U = I_q$ to express $W^\star$ in terms of U:

$$\begin{aligned} W^\star = \Sigma_Y(\Sigma_Y + \Sigma_N)^{-1} &= (UDU^T)\ \left(UDU^T + \sigma_N^2 UU^T\right)^{-1} \\ &= (UDU^T)\, U \left(D + \sigma_N^2 I_n\right)^{-1} U^T \\ &= U \left[\frac{\sigma_j^2}{\sigma_j^2 + \sigma_N^2}\right] U^T. \end{aligned} \tag{12.11}$$

Note that for each $j = 1, \ldots, q$,

$$0 \leq \frac{\sigma_j^2}{\sigma_j^2 + \sigma_N^2} \leq 1.$$

So $W^\star$ selectively attenuates the deviations of X from its mean in the direction of each of the eigenvectors u_j, $j = 1, \ldots, q$. For eigenvectors u_j with $\sigma_j^2 \gg \sigma_N^2$, $\sigma_j^2/(\sigma_j^2 + \sigma_N^2) \approx 1$. Hence $W^\star$ has approximately unity gain in the direction u_j. In contrast, if $\sigma_N^2 \gg \sigma_j^2$, then $\sigma_j^2/(\sigma_j^2 + \sigma_N^2) \ll 1$ and $W^\star$ has very low gain in the direction u_j. In such directions $X - \mu_X$ is highly attenuated. In situations between these extremes, the amount of attenuation in the direction u_j is determined by the ratio $\sigma_j^2/(\sigma_j^2 + \sigma_N^2)$.

Now bring the mean μ_Y into the picture by using (12.11) to rewrite (12.2) in the form

$$\hat{y} = U \left(\left[\frac{\sigma_j^2}{\sigma_j^2 + \sigma_N^2}\right] U^T x + \left[\frac{\sigma_W^2}{\sigma_j^2 + \sigma_N^2}\right] U^T \mu_Y \right). \tag{12.12}$$

Notice that $\alpha_j = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_N^2}$ and $\beta_j = \frac{\sigma_N^2}{\sigma_j^2 + \sigma_N^2}$ sum to 1. So α_j small means β_j must be large to ensure the sum is 1. Hence if $\sigma_j^2 < \sigma_N^2$, α_j is small and the component of $\hat{y}$ in the direction u_j is formed mainly from the component of μ_Y in this direction. Conversely, when $\sigma_j^2 > \sigma_N^2$, α_j is large and the component of $\hat{y}$ in the direction u_j is formed mainly from the component of x in this direction. In particular, when $\sigma_N^2 \gg \sigma_j$, for all $j = 1, \ldots, q$, the optimal affine estimate of Y given $X = x$ reverts to the mean μ_Y. This makes sense since the observation $X = x$ adds very little information about the value of Y. Hence the best we can do under the MSE metric is estimate the value of Y to be its mean μ_Y.

12.4.2 Special Case: Estimating a Latent Vector from Linear, Noisy Observations

We now consider a slightly more complex situation. Let the random vector Y model labels of interest and assume that Y has known mean μ_Y and covariance Σ_Y. Let the random vector N represent noise and assume that N has zero mean, known covariance Σ_N, and that Y and N are independent.

Here is the new twist. Let $B \in \mathbb{R}^{n \times p}$ be given matrix and assume that the examples are linear functions of the labels plus noise:

$$X = BY + N. \tag{12.13}$$

Given $X = x$ and knowledge of B, μ_Y, Σ_Y and Σ_N we want to estimate the value of the label vector Y.

At first one might try multiplying both sides of the above equation by B^{-1} but there is no guarantee that B is invertible and even if it is, it may have small eigenvalues giving B^{-1} a large gain in some directions.

Such large gains could significantly amplify the noise N. An alternative is to use the pseudoinverse B^+ of B to form $\hat{y} = B^+x$. This takes care of the zero eigenvalues of B, but this could still exhibit large gains if B has small nonzero eigenvalues. Since the noise is the source of the problem, one might consider first denoising to estimate BY, then multiplying that result by B^+. Let's come back to that idea at the end.

We proceed as before and seek the best affine estimate $\hat{y}$ of the value of Y given $X = x$. We have

$$\begin{aligned}
\mu_X &= E[BY + N] = B\mu_Y \\
\Sigma_X &= E[B(Y - \mu_Y)(Y - \mu_Y)^T B^T + NN^T] = B\Sigma_Y B^T + \Sigma_N \\
\Sigma_{YX} &= E[(Y - \mu_Y)(B(Y - \mu_Y) + N)^T] = \Sigma_Y B^T.
\end{aligned}$$

Hence the minimum MSE affine estimator is given by

$$\begin{aligned}
W^\star &= \Sigma_Y B^T (B\Sigma_Y B^T + \Sigma_N)^{-1} \\
\hat{y}(x) &= W^\star(x - \mu_X) + \mu_Y.
\end{aligned} \tag{12.14}$$

12.5 Appendix: Exchanging Differentiation with Expectation

This appendix covers advanced material intended for those who have an interest in the fine details. In showing that the mean (or conditional mean) was an optimal MSE estimator we exchanged differentiation with an expectation (an integral). In simplest terms, we have a real valued function $f(y, a)$, where $y, a \in \mathbb{R}^n$, and want to execute the exchange:

$$D_a E[f(Y, a)](h) = E[D_a f(Y, a)(h)].$$

Differentiation is a limit operation:

$$\lim_{\alpha \to 0} \frac{f(y, a + \alpha h) - f(y, a)}{\alpha}.$$

So we want to exchange the order of a limit operation and an integral (expectation). The dominated convergence theorem gives sufficient conditions under which this is possible.

In Lemma 12.2.1, the real valued function is $f(y, a) = \|y - a\|_2^2$. Assume μ_Y is finite and restrict a to lie in an open ball B around μ_Y. In addition, restrict h so that $a + h \in B$. If there exists a nonnegative function $g(y)$ such that $E[g(Y)]$ is finite and for α sufficiently small:

$$\left| \frac{f(y, a + \alpha h) - f(y, a)}{\alpha} \right| \leq g(y),$$

then the exchange of the derivative and expectation is valid. To demonstrate such a function we proceed as follows:

$$\begin{aligned}
\left| \frac{f(y, a + \alpha h) - f(y, a)}{\alpha} \right| &= |D_a f(y, a + \theta h)h| \,, \text{ some } \theta \in (0, \alpha) \text{ (Mean Value Theorem)} \\
&\leq \|\nabla f(y, a')\|_2 \|h\|_2, \; a' = a + \theta h \text{ (Cauchy-Schwartz)} \\
&\leq C\|\nabla f(y, a')\|_2 \; (h \text{ is bounded}) \\
&= 2C\|y - a'\|_2, \text{ (by defn of } f(y, a) \text{)} \\
&\leq 2C(\|y - \mu_Y\|_2 + \|\mu_Y - a'\|_2) \\
&= 2C\|y - \mu_Y\|_2 + C'.
\end{aligned}$$

The final step is stated as the following lemma.

Lemma 12.5.1. Let the random vector Y have finite mean and finite covariance. Then

$$E[\|Y - \mu_Y\|_2] \leq 1 + \text{trace}(\Sigma_Y). \tag{12.15}$$

Proof. Let $\mathbb{I}(y)$ be the indicator function of the set $\{y\colon \|y - \mu_Y\|_2 < 1\}$. Then

$$\begin{aligned}\|y - \mu_Y\|_2 &= \|y - \mu_Y\|_2\mathbb{I}(y) + \|y - \mu_Y\|_2(1 - \mathbb{I}(y)) \\ &\leq 1\mathbb{I}(y) + \|y - \mu_Y\|_2^2(1 - \mathbb{I}(y)) \\ &\leq 1 + \|y - \mu_Y\|_2^2.\end{aligned}$$

Thus $E[\|Y - \mu_Y\|_2] \leq 1 + \text{trace}(\Sigma_Y)$. □

12.6 Exercises

Exercise 12.1. Show that the MSE of the estimator (12.2) is $\text{trace}\left(\Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\right)$.

Exercise 12.2. Show that the MSE of the estimator (12.10) is $\text{trace}\left((I - W^\star)\Sigma_Y\right)$.

Exercise 12.3. Show that the MSE of the estimator (12.14) is $\text{trace}\left((I - W^\star B)\Sigma_Y\right)$.

Exercise 12.4. Show that using minimum MSE affine denoising to estimate BY given $X = x$, then multiplying the result by B^+, yields the estimator in (12.14).