Machine learning notes

Francesco Boi

1 Preliminary definitions

1.1 Trace of a matrix

The **trace** of a square matrix $\mathbf{A}$, denoted as $\mathbf{Tr}(\mathbf{A})$ is the sum of diagonal elements:

$$\mathbf{Tr}(\mathbf{A}) = \sum_{d=1}^{p} A_{dd} \tag{1}$$

It follows that $\mathrm{Tr}(\mathbf{I_d}) = p$. Also $\mathbf{Tr}(\mathbf{AB}) = \mathbf{Tr}(\mathbf{AB})$ and $\mathbf{Tr}(\mathbf{x}^T\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ the latter being a scalar.

1.2 Expectation

Definition 1.1. Expectation Let X be a random variable with a finite number of outcomes $x_1, x_2, \cdots, x_k$ occurring respectively with probabilities $p_1, p_2, \cdots, p_k$. The expectation value is the summation of each outcome times its probability.

$$\mathbf{E}[X] = \sum_{k} x_k \cdot p_k \tag{2}$$

In case of an infinite number of outcomes the summation is replaced with the integral:

$$\mathbf{E}[X] = \int x \cdot p(x)\, dx \tag{3}$$

As explained here, when many random variables are involved, and there is no subscript in the E symbol, the expected value is taken with respect to their joint distribution:

$$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{XY}(x,y)\, dx\, dy$$

When a subscript is present, in some cases it tells us on which variable we should condition. So

$$\mathrm{E}_{\mathrm{X}}[\mathrm{h}(\mathrm{X},\mathrm{Y})] = \mathrm{E}[\mathrm{h}(\mathrm{X},\mathrm{Y}) \mid \mathrm{X}] = \int_{-\infty}^{\infty} \mathrm{h}(\mathrm{x},\mathrm{y}) \mathrm{f}_{\mathrm{h}(\mathrm{X},\mathrm{Y})|\mathrm{X}}(\mathrm{h}(\mathrm{x},\mathrm{y}) \mid \mathrm{x})\, \mathrm{dh}$$

...But in other cases, it tells us which density to use for the "averaging"

$$\mathrm{E}_{\mathrm{X}}[\mathrm{h}(\mathrm{X},\mathrm{Y})] = \int_{-\infty}^{\infty} \mathrm{h}(\mathrm{x},\mathrm{y}) \mathrm{f}_{\mathrm{X}}(\mathrm{x})\, \mathrm{dx}$$

1.3 Variance

Definition 1.2. Variance The variance of a random variable X is the expected value of the squared deviation from the mean of X:

$$\operatorname{Var}(\mathrm{X}) = \mathrm{E}\left[(\mathrm{X} - \mu)^2\right] \tag{4}$$

Variance can be expressed in another way recalling $\mu = \mathrm{E}[\mathrm{X}]$ and using the linearity property:

$$\begin{aligned} \operatorname{Var}(\mathrm{X}) &= \mathrm{E}\left[(\mathrm{X} - \mu)^2\right] = \mathrm{E}\left[(\mathrm{X} - \mathrm{E}[\mathrm{X}])^2\right] = \\ &= \mathrm{E}\left[\mathrm{X}^2 - 2 \cdot \mathrm{X} \cdot \mathrm{E}[\mathrm{X}] + \mathrm{E}[\mathrm{X}]^2\right] \\ &= \mathrm{E}\left[\mathrm{X}^2\right] - 2 \cdot \mathrm{E}[\mathrm{X} \cdot \mu] + \mathrm{E}\left[\mu^2\right] \\ &= \mathrm{E}\left[\mathrm{X}^2\right] - 2 \cdot \mu \cdot \mathrm{E}[\mathrm{X}] + \mu^2 \\ &= \mathrm{E}\left[\mathrm{X}^2\right] - \mu^2 = \mathrm{E}\left[\mathrm{X}^2\right] - \mathrm{E}[\mathrm{X}]^2 \end{aligned} \tag{5}$$

1.4 Median

Definition 1.3. Median For any probability distribution on the real line $\mathbb{R}$ with cumulative distribution function F, regardless of whether it is any kind of continuous probability distribution, in particular an absolutely continuous distribution (which has a probability density function), or a discrete probability distribution, a median is by definition any real number m that satisfies

the inequalities:

$$P(x \leq m) \geq \frac{1}{2} \quad \text{and} \quad P(x \leq m) \geq \frac{1}{2} \tag{6}$$

$$\tag{7}$$

or equivalently the inequalities

$$\int_{-\infty}^{m} F(x)dx \geq \frac{1}{2} \quad \text{and} \quad \int_{m}^{\infty} F(x)dx \geq \frac{1}{2} \tag{8}$$

1.5 Median as the minimizer of L_1 norm

Assume that S is a finite set, with say k elements. Line them up in order, as $s_1 < s_2 < \cdots < s_k$.

If k is even there are (depending on the exact definition of median) many medians. $|x-s_i|$ is the *distance* between x and s_i, so we are trying to minimize the sum of the distances. For example, we have k people who live at various points on the x-axis. We want to find the point(s) x such that the sum of the travel distances of the k people to x is a minimum.

Imagine that the s_i are points on the x-axis. For clarity, take $k = 7$. Start from well to the left of all the s_i, and take a tiny step, say of length ϵ, to the right. Then you have gotten ϵ closer to every one of the s_i, so the sum of the distances has decreased by 7ϵ.

Keep taking tiny steps to the right, each time getting a decrease of 7ϵ. This continues until you hit s_1. If you now take a tiny step to the right, then your distance from s_1 increases by ϵ, and your distance from each of the remaining s_i decreases by ϵ. So there is a decrease of 6ϵ, and an increase of ϵ, for a net decrease of 5ϵ in the sum.

This continues until you hit s_2. Now, when you take a tiny step to the right, your distance from each of s_1 and s_2 increases by ϵ, and your distance from each of the five others decreases by ϵ, for a net decrease of 3ϵ.

This continues until you hit s_3. The next tiny step gives an increase of 3ϵ, and a decrease of 4ϵ, for a net decrease of ϵ.

This continues until you hit s_4. The next little step brings a total increase of 4ϵ, and a total decrease of 3ϵ, for an increase of ϵ. Things get even worse when you travel further to the right. So the minimum sum of distances is reached at s_4, the median.

The situation is quite similar if k is even, say $k = 6$. As you travel to the right, there is a net decrease at every step, until you hit s_3. When you

are between s_3 and s_4, a tiny step of ϵ increases your distance from each of s_1, s_2, and s_3 by ϵ. But it decreases your distance from each of the three others, for no net gain. Thus any x in the interval from s_3 to s_4, including the endpoints, minimizes the sum of the distances.

In the even case, Some people prefer to say that any point between the two "middle" points is a median. So the conclusion is that the points that minimize the sum are the medians. Other people prefer to define the median in the even case to be the average of the two "middle" points. Then the median does minimize the sum of the distances, but some other points also do.

In formulas consider two x_i's x_1 and x_2, with $x_2 > x_1$

- $$x_1 \leq a \leq x_2$$
$$\sum_{i=1}^{2} |x_i - a| = |x_1 - a| + |x_2 - a| = a - x_1 + x_2 - a = x_2 - x_1 \tag{9}$$

- $$a < x_1$$
$$\sum_{i=1}^{2} |x_i - a| = x_1 - a + x_2 - a = x_1 + x_2 - 2a \geq x_1 + x_2 - 2x_1 \tag{10}$$
$$= x_2 - x_1$$

- $$a \geq x_2$$
$$\sum_{i=1}^{2} |x_i - a| \quad = -x_1 + a - x_2 + a = -x_1 - x_2 + 2a \geq -x_1 - x_2 + 2x_2 =$$
$$= x_2 - x_1 \tag{11}$$

$\Longrightarrow$ for any two x_i's the sum of the absolute values of the deviations is minimum when $x_1 \leq a \leq x_2$ or $a \in [x_1, x_2]$.

When n is odd,

$$\sum_{i=1}^{n} |x_i - a| = |x_1 - a| + |x_2 - a| + \cdots + \left|x_{\frac{n-1}{2}} - a\right| + \left|x_{\frac{n+1}{2}} - a\right| + \\ + \left|x_{\frac{n+3}{2}} - a\right| + \cdots + |x_{n-1} - a| + |x_n - a| \tag{12}$$

consider the intervals $[x_1, x_n], [x_2, x_{n-1}], [x_3, x_{n-2}], \ldots, \left[x_{\frac{n-1}{2}}, x_{\frac{n+3}{2}}\right]$. If a is a member of all these intervals. i.e, $\left[x_{\frac{n-1}{2}}, x_{\frac{n+3}{2}}\right]$,

using the above theorem, we can say that all the terms in the sum except $\left|x_{\frac{n+1}{2}} - a\right|$ are minimized. So

$$\sum_{i=1}^{n} |x_i - a| = (x_n - x_1) + (x_{n-1} - x_2) + (x_{n-2} - x_3) + \cdots + \\ + \left(x_{\frac{n+3}{2}} - x_{\frac{n-1}{2}}\right) + \left|x_{\frac{n+1}{2}} - a\right| = \left|x_{\frac{n+1}{2}} - a\right| + \text{costant} \tag{13}$$

To minimize also the term $\left|x_{\frac{n+1}{2}} - a\right|$ it is clear we have to choose $a = x_{\frac{n+1}{2}}$ to get 0 but this is the definition of the median.

$\Longrightarrow$ When n is odd,the median minimizes the sum of absolute values of the deviations.

When n is even,

$$\sum_{i=1}^{n} |x_i - a| = |x_1 - a| + |x_2 - a| + \cdots + |x_{\frac{n}{2}} - a| + \\ + |x_{\frac{n}{2}+1} - a| + \cdots + |x_{n-1} - a| + |x_n - a| \tag{14}$$

If a is a member of all the intervals $[x_1, x_n], [x_2, x_{n-1}], [x_3, x_{n-2}], \ldots, \left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right]$, i.e, $a \in \left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right]$,

$$\sum_{i=1}^{n} |x_i - a| = (x_n - x_1) + (x_{n-1} - x_2) + (x_{n-2} - x_3) + \cdots + \\ + \left(x_{\frac{n}{2}+1} - x_{\frac{n}{2}} \right) \tag{15}$$

$\implies$ When n is even, any number in the interval $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$, i.e, including the median, minimizes the sum of absolute values of the deviations. For example consider the series:$2, 4, 5, 10$, median, $M = 4.5$.

$$\sum_{i=1}^{4} |x_i - M| = 2.5 + 0.5 + 0.5 + 5.5 = 9$$

If you take any other value in the interval $\left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right] = [4, 5]$, say 4.1

$$\sum_{i=1}^{4} |x_i - 4.1| = 2.1 + 0.1 + 0.9 + 5.9 = 9$$

Taking for example 4 or 5 yields the same result:

$$\sum_{i=1}^{4} |x_i - 4| = 2 + 0 + 1 + 6 = 9$$

$$\sum_{i=1}^{4} |x_i - 5| = 3 + 1 + 0 + 5 = 9$$

This is because when summing the distance from a to the two middle points, you end up with the distance between them: $a - x_{\frac{n}{2}} + (x_{\frac{n}{2}+1} - a) = x_{\frac{n}{2}+1} - x_{\frac{n}{2}}$

For any value outside the interval $\left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right] = [4, 5]$, say 5.2

$$\sum_{i=1}^{4} |x_i - 5.2| = 3.2 + 1.2 + 0.2 + 4.8 = 9.4$$

1.6 Gaussian function and gaussian distribution

Definition 1.4. Gaussian function A Gaussian function is a mathematical function in the form:

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}} \tag{16}$$

The Gaussian function has three parameters a,b and c. The graph of a Gaussian function is the Bell curve. The parameter a is the height of the curve's peak, b is the position of the peak and c the *standard deviation* controls the width of the bell.

An important property of the Gaussian function is that the product of two Gaussian functions is still a Gaussian function:

$$\begin{aligned}
f(x)\cdot g(x) &= ae^{-\frac{(x-b)^2}{2c^2}} a_1 e^{-\frac{(x-b_1)^2}{2c_1^2}} = (a\cdot a_1)e^{-\frac{x^2-2bx+b^2+x^2-2b_1x+b_1^2}{2c^2 2c_1^2}} = \\
&= (a\cdot a_1)e^{-\frac{2x^2-2x(b+b_1)+b^2+b_1^2}{2c^2 2c_1^2}} = (a\cdot a_1)e^{-\frac{x^2-2x\frac{b+b_1}{2}+\frac{b^2+b_1^2}{2}}{2c^2c_1^2}} = \\
&= (a\cdot a_1)e^{-\frac{x^2-2x\frac{b+b_1}{2}+\frac{b^2+b_1^2}{2}}{2c^2c_1^2}} e^{\frac{\left(\frac{b+b_1}{2}\right)^2-\left(\frac{b+b_1}{2}\right)^2}{2c^2c_1^2}} = \\
&= (a\cdot a_1)e^{-\frac{x^2-2x\frac{b+b_1}{2}+\left(\frac{b+b_1}{2}\right)^2}{2c^2c_1^2}} e^{\frac{\left(\frac{b+b_1}{2}\right)^2-\frac{b^2+b_1^2}{2}}{2c^2c_1^2}} = \\
&= (a\cdot a_1)e^{-\frac{\left(x-\frac{b+b_1}{2}\right)^2}{2c^2c_1^2}} e^{\frac{\left(\frac{b+b_1}{2}\right)^2-\frac{b^2+b_1^2}{2}}{2c^2c_1^2}} = \\
&= a_2 e^{-\frac{\left(x-\frac{b+b_1}{2}\right)^2}{2c^2c_1^2}} = a_2 e^{-\frac{(x-b_2)^2}{2c_2^2}}
\end{aligned} \tag{17}$$

Definition 1.5. Gaussian distribution A normalized Guassian function or normal distribution is a Gaussian function with an area under the curve equal to 1. Hence it can be interpreted as a probability distribution.

To find the formula let us force the curve to have area 1:

$$\int_{-\infty}^{\infty} ae^{-\frac{(y-b)^2}{2c^2}}\,dy = a\int_{-\infty}^{\infty} e^{-\frac{(y-b)^2}{2c^2}}\,dy \tag{18}$$

and performing a change of integration variable:

$$\begin{aligned} &\frac{y-b}{\sqrt{2}c} = x \Rightarrow dx = dy\frac{1}{\sqrt{2}c}dx \Rightarrow dy = \sqrt{2}cdx \\ &a\int_{-\infty}^{\infty} e^{-\frac{(y-b)^2}{2c^2}}dy = \sqrt{2}ac\int_{-\infty}^{\infty} e^{-x^2}dx \end{aligned} \tag{19}$$

where $\sqrt{2}c$ is a constant so that in can be moved out of the integral. From now on we will focus just on the integral.

Unfortunately the integral has not solutions with elementary functions (Louiville theorem, see Abstract Algebra). However the definite integral exists (demonstration is skipped) and it is:

$$\int_{-\infty}^{\infty} e^{-x^2}dx = \sqrt{\pi} \tag{20}$$

To have area equal to 1 we must have:

$$ac\sqrt{2\pi} = 1 \Rightarrow a = \frac{1}{\sqrt{2\pi}c} \tag{21}$$

So now the function is defined by just 2 parameters.

Normally $b = \mu$ and is the mean of the function while $c = \sigma$ is the standard deviation.

As opposed to the Gaussian function, multiplying two of these functions we do not get another gaussian distribution, but a scaled guassian distribution. If the Gaussian functions have 0-mean and $\sigma^2 = 1$ then the product is still Guassian distribution with 0-mean and $\sigma^2 = 1$.

Definition 1.6. Standard Normal Gaussian distribution A Standard Normal Gaussian distribution is a Gaussian distribution with 0 mean and standard deviation 1.

2 Statistical Decision Theory

2.1 Expected prediction error

Let $X \in \mathbb{R}^p$ denote a real valued random input vector, and $Y \in \mathbb{R}^p$ a real valued random output variable, with joint distribution $Pr(X, Y)$. We seek

a function f(X) for predicting Y given values of the input X. This theory requires a loss function $L(Y, f(X))$ for penalizing errors in prediction, and by far the most common and convenient is squared error loss: $L(Y, f(X)) = (Y - f(X))^2$.

The estimated prediction error is

$$\begin{aligned} \mathrm{EPE}(f) &= \mathbf{E}\left[(Y - f(X))^2\right] = \int (y - f(x))^2 \, p(x, y) dx dy \\ &= \int_x \int_y (y - f(x))^2 \, p(x, y) dx dy \end{aligned} \tag{22}$$

Recalling $p(x, y) = p(y|x)p(x)$:

$$\begin{aligned} \mathrm{EPE}(f) &= \int_x \int_y (y - f(x))^2 \, p(y|x)p(x) dx dy \\ &= \int_x \int_y \left((y - f(x))^2 \, p(y|x) dy\right) p(x) dx \\ &= \int_x \mathbf{E}_{Y|X}\left[(y - f(X))^2 \, |X = x\right] p(x) dx \\ &= \mathbf{E}_X\left[\mathbf{E}_{Y|X}\left[(y - f(X))^2 \, |X = x\right]\right] \end{aligned} \tag{23}$$

So to minimize the prediction error:

$$f(x) = \arg\min_c \mathbf{E}_{Y|X}\left[(y - c(x))^2 \, |X = x\right] \quad \Rightarrow f(x) = \mathbf{E}\left[Y|X = x\right] \tag{24}$$

The *Nearest Neighbour* algorithm, assigns labels to points by counting and averaging the labels of the points belonging to a given neighbourhood:

$$\hat{f}(x) = \mathrm{Ave}\left(y_i | x_i \in N_k(x)\right) \tag{25}$$

where $N_k(x)$ contains the k points closest to x. This presents **two approximations**

- expectation is approximated by averaging
- conditioning at a point is relaxed to conditioning on some region centred at the target point.

With k sufficiently large, the average gets more stable and with large N the points will be more likely close to x. If $k, N \to \infty$ with $k/N \to 0$ the average becomes the expectation and we have the best possible estimator.

Unfortunately often we do not have so much data and some other times we might want exploit the supposed structure of data (linear, polynomial etc.).

However there is even a bigger problem when there are too many dimensions (i.e., p is large). Consider a uniformly distributed input in a p dimensional unit hypercube. Consider a hypercube neighbourhood around the target point capturing a fraction r of the total observations distributed among the unit hypercube. The edge of the neighbour hypercube will be $e_p(r) = r^{1/p}$. In 10 dimensions, using a neighborhood capturing 1% of the data we have $r(0.01) = 0.63$ so we must use 63% of the total data for one target.

On the contrary the linear regression is a **model-based approach**, i.e., one assumes that the function f(x) is approximately linear:

$$f(x) \approx x^T \beta \tag{26}$$

Putting this in the EPE equation:

$$\begin{aligned} f(x) &= \arg\min_c \mathbf{E}_{Y|X} \left[(y - c(x))^2 \, | X = x \right] \\ &= \arg\min_\beta \mathbf{E}_{Y|X} \left[\left(y - x^T \beta \right)^2 | X = x \right] \\ \Rightarrow \beta &= \left[\mathbf{E} \left[X \cdot X^T \right] \right]^{-1} \cdot \mathbf{E} \left[X \cdot Y \right] \end{aligned} \tag{27}$$

The minimum of a quadratic function is given by deriving and setting its derivative to 0.

If instead of a L_2 loss function we use L_1

$$\begin{aligned} \mathrm{EPE}(f) &= \mathbf{E} \left[|Y - f(X)| \right] = \int |y - f(x)| \, p(x, y) dx dy \\ &= \int_x \int_y |y - f(x)| \, p(x, y) dx dy \end{aligned} \tag{28}$$

Recalling $p(x, y) = p(y|x)p(x)$:

$$\begin{aligned} \text{EPE}(f) &= \int_x \int_y |y - f(x)|\, p(y|x)p(x)dxdy \\ &= \int_x \int_y \left(|y - f(x)|\, p(y|x)dy\right) p(x)dx \\ &= \int_x \mathbf{E}_{Y|X}\left[|y - f(X)|\, |X = x\right] p(x)dx \\ &= \mathbf{E}_X \left[\mathbf{E}_{Y|X}\left[|y - f(X)|\, |X = x\right]\right] \end{aligned} \tag{29}$$

$$f(x) = \arg\min_c \mathbf{E}_{Y|X}\left[|y - c(x)|\, |X = x\right] \tag{30}$$

and as already seen in 1.5, the minimizer for the sum of distances is the median.

2.1.1 Loss function for categorical variables

For categorical output variables $\mathbb{G}_k$ we have:

$$\text{EPE} = \mathbf{E}\left[L\left(G, \hat{G}(X)\right)\right] = \mathbf{E}_x \sum_{k=1}^{K} L\left[\mathbb{G}_k, \hat{G}(X)\right] \Pr\left(\mathbb{G}_k|X\right) \tag{31}$$

where the expectation is again taken with respect to the joint distribution $\Pr(G, X)$. Conditioning again we can write:

$$\text{EPE} = \mathbf{E}_x \sum_{k=1}^{K} L\left[\mathbb{G}_k, \hat{G}(X)\right] \Pr\left(\mathbb{G}_k|X\right) \tag{32}$$

where the integral over y has been substituted with the summation due to the categorical nature of the variable.

The minimizer is given by:

$$\hat{G}(x) = \arg\min_{g \in \mathbb{G}} \sum_{k=1}^{K} L\left(\mathbb{G}_k, g\right) \Pr\left(\mathbb{G}_k|X = x\right) \tag{33}$$

Often the *zero-one loss function* is used for categorical variables and the above simplifies to:

$$\hat{G}(x) = \arg\min_{g \in \mathbb{G}} \left[1 - \Pr\left(g|X = x\right)\right] = \arg\max_{g \in \mathbb{G}} \left[\Pr\left(g|X = x\right)\right] \tag{34}$$

This is known as *Bayes classifier* because **it classifies to the most probable class, using conditional discrete probability distribution**.

Note: When models or loss functions use additional parameters that penalize complexity (Lasso, Ridge and others) we cannot use the training data to determine these parameters, since we would pick those that gave interpolating fits with zero residuals but it will be unlikely to predict future data.

2.2 Bias-Variance trade-off

$$\begin{aligned} \mathrm{EPE(f)} &= \mathbf{E}\left[(\mathrm{Y} - \mathrm{f(X)})^2\right] = \\ &= \mathbf{E}\left[\mathrm{Y}^2\right] - 2\mathbf{E}\left[\mathrm{Y}\right]\mathbf{E}\left[\mathrm{f(X)}\right] + \mathbf{E}\left[\mathrm{f(X)}^2\right] \\ &= \mathrm{Y}^2 - 2\mathrm{Y}\mathbf{E}\left[\mathrm{f(X)}\right] + \mathbf{E}\left[\mathrm{f(X)}^2\right] \end{aligned} \tag{35}$$

Recalling

$$\begin{aligned} &\mathrm{BIAS}\left(\mathrm{Y}, \mathbf{E}\left[\mathrm{f(X)}\right]\right) = \left|\mathrm{Y} - \mathbf{E}\left[\mathrm{f(X)}\right]\right| \\ &\Rightarrow \mathrm{BIAS}\left(\mathrm{Y}, \mathrm{f(X)}\right)^2 = \left(\mathrm{Y} - \mathbf{E}\left[\mathrm{f(X)}\right]\right)^2 \\ &= \mathrm{Y}^2 - 2\mathrm{Y}\mathbf{E}\left[\mathrm{f(X)}\right] + \mathbf{E}\left[\mathrm{f(X)}\right]^2 \end{aligned} \tag{36}$$

EPE can be expressed using also the Variance definition in 1.3

$$\begin{aligned} \mathrm{EPE(f)} &= \mathrm{Y}^2 - 2\mathrm{Y}\mathbf{E}\left[\mathrm{f(X)}\right] + \mathbf{E}\left[\mathrm{f(X)}^2\right] + \mathbf{E}\left[\mathrm{f(X)}\right]^2 - \mathbf{E}\left[\mathrm{f(X)}\right]^2 \\ &= \mathrm{BIAS}(\mathrm{Y}, \mathrm{f(X)}^2 + \mathrm{Var}(\mathrm{Y}, \mathrm{f(x)}) \end{aligned} \tag{37}$$

The bias is given by the distance of our predictions from real points. Complex models have more degrees of freedom and are able to fit closer real points hence they to tend to have low bias. However, they present higher variance. On the contrary simple models (i.e., linear) have lower variance but higher bias.

The error due to variance is the amount by which the prediction, over one training set, differs from the expected predicted value, over all the training sets. As with bias, you can repeat the entire model building process multiple times. To paraphrase Manning et al (2008), variance measures how inconsistent are the predictions from one another, over different training sets, not whether they are accurate or not. A learning algorithm with low bias must

be "flexible" so that it can fit the data well. But if the learning algorithm is too flexible, it will fit each training data set differently, and hence have high variance.

3 Linear Regression Models

We start from the *univariate linear regression*, i.e., each output consists of a single value while the input is a vector of values.

3.1 Univariate linear regression

Univariate means single output, i.e, y is a number. The basic form is

$$\mathrm{f(X)} = \beta_0 + \sum_{\mathrm{j–1}}^{\mathrm{p}} \mathrm{X_j}\beta_\mathrm{j} \tag{38}$$

The most popular estimation method for a linear model is the least square:

$$\mathrm{RSS}(\beta) = \sum_{\mathrm{i=1}}^{\mathrm{p}} (\mathrm{y_i} - \mathrm{f(x_i)})^2 = (\mathbf{y} - \mathbf{X}\beta)^{\mathrm{T}} (\mathbf{y} - \mathbf{X}\beta) \tag{39}$$

where $\mathbf{X}$ is a $\mathrm{N} \times (\mathrm{p}+1)$ matrix with each row being an input vector, $\mathbf{y}$ a N vector (we must have N input-output pairs, in this case the output is considered mono-dimensional).

By minimizing we get:

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathrm{Y} \tag{40}$$

Geometrically we are projecting y onto the hyperplane spanned by X and the projections is referred to as $\hat{\mathrm{y}}$:

$$\hat{\mathrm{y}} = \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\mathrm{Y} = \mathrm{HY} \tag{41}$$

where H is called the *hat* matrix.

3.2 Equivalence of Ordinary least squares and maximum likelihood

We are using an additive model, assuming a Gaussian white noise:

$$\mathrm{y} = \beta^{\mathrm{T}}\mathbf{x} + \epsilon \tag{42}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{43}$$

Adding a constant to a Gaussian random variable is equivalent to another Gaussian random variable with the mean shifted:

$$\Pr(\mathrm{y}) \sim \mathcal{N}(\beta^{\mathrm{T}}\mathbf{x}, \sigma^2) \tag{44}$$

Considering the matrix $\mathbf{X}$ and the output vector $\mathbf{y}$ representing the training set, used to estimate the coefficients, we have:

$$\Pr(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \Pr(\mathrm{y_i}|\mathbf{x}, \beta, \sigma^2) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathcal{N}(\beta^{\mathrm{T}}\mathbf{x}, \sigma^2) \tag{45}$$

where we have assumed each observation is independent. A product of univariate Gaussian can be rewritten as a multivariate Gaussian:

$$\begin{aligned}
\Pr(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) &= \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathcal{N}(\beta^{\mathrm{T}}\mathbf{x}, \sigma^2) = \prod_{\mathrm{i}=1}^{\mathrm{N}} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \mathrm{e}^{-\frac{\left(\mathrm{y_i}-\beta^{\mathrm{T}}\mathbf{x_i}\right)^2}{2\sigma^2}} = \\
&= \frac{1}{(2\pi)^{\frac{\mathrm{p}}{2}}\sigma} \prod_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{e}^{-\sum_{\mathrm{i}=1}^{\mathrm{N}} \frac{\left(\mathrm{y_i}-\beta^{\mathrm{T}}\mathbf{x_i}\right)^2}{2\sigma^2}} = \\
&= \frac{1}{(2\pi)^{\frac{\mathrm{p}}{2}}\sigma|\mathbf{I}|} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})} = \mathcal{N}(\beta^{\mathrm{T}}\mathbf{X}, \sigma^2\mathbf{I})
\end{aligned} \tag{46}$$

If the variables are not independent the more general form is:

$$\mathcal{N}(\beta^{\mathrm{T}}\mathbf{X}, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{\mathrm{p}}{2}}|\Sigma|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})} \tag{47}$$

Definition of likelihood The quantity $\Pr(\mathrm{y}|\mathbf{x}, \beta, \sigma^2)$ is called **likelihood** and tell us how much it is likely the outcome $\mathrm{y_i}$ in the dataset given the input $\mathbf{x}$ and the parameters.

A different approach to find a model that fits the data is to maximize the *likelihood* of the whole dataset:

$$\mathrm{L} = \Pr(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) = \frac{1}{(2\pi)^{\frac{\mathrm{p}}{2}}|\Sigma|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}-\beta^{\mathrm{T}}\mathbf{X})} \tag{48}$$

Actually maximizing the likelihood is equivalent to maximizing its logarithmic, the *log-likelihood*:

$$\begin{aligned}
\log \mathrm{L} &= \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \left[\frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \mathrm{e}^{-\frac{\left(\mathrm{y_i}-\beta^{\mathrm{t}}\mathbf{x_i}\right)^2}{2\sigma^2}} \right] \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} -\frac{1}{2}\log 2\pi - \log\sigma - \frac{\left(\mathrm{y_i} - \beta^{\mathrm{T}}\mathbf{x_i}\right)^2}{2\sigma^2} = \\
&= -\frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2\sigma^2}\sum_{\mathrm{i}=1}^{\mathrm{N}} \left(\mathrm{y_i} - \beta^{\mathrm{T}}\mathbf{x_i}\right)^2
\end{aligned} \tag{49}$$

As already done for OLS, taking the derivative w.r.t. β and setting it to 0:

$$\begin{aligned}
&\frac{\partial \log \mathrm{L}}{\partial \beta} = -\frac{1}{2\sigma^2}(-2)\left(\mathbf{y_i} - \beta^{\mathrm{T}}\mathbf{x_i}\right) = 0 \\
&\Rightarrow \left(\mathbf{y} - \beta^{\mathrm{T}}\mathbf{X}\right) = 0 \Rightarrow \beta = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}
\end{aligned} \tag{50}$$

This is the same solution of the OLS: the two models are equivalent.
The two models are equivalent assuming a normal distribution.

3.3 Expectation of the parameter estimation: unbiased estimator

Computing the expectation of $\hat{\beta}$:

$$\begin{aligned}
&\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\beta}\right] = \sum \hat{\beta}\Pr(\mathbf{y}|\mathrm{X}, \beta, \sigma^2) = \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \sum \mathbf{y}\Pr(\mathbf{y}|\mathrm{X}, \beta, \sigma^2) = \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\right] = \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta = \beta
\end{aligned} \tag{51}$$

This is an **unbiased estimator**.

Now let us calculate the covariance matrix:

$$\begin{aligned}
&\mathrm{Cov}\left[\hat{\beta}\right] = \\
&= \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\left(\hat{\beta} - \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\beta\right]\right)\left(\hat{\beta} - \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\beta\right]\right)^{\mathrm{T}}\right] = \\
&= \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^{\mathrm{T}}\right] = \\
&= \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\beta}\hat{\beta}^{\mathrm{T}}\right] - \beta\beta^{\mathrm{T}}
\end{aligned} \tag{52}$$

and

$$\begin{aligned}
&\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\beta}\hat{\beta}^{\mathrm{T}}\right] = \\
&= \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)^{\mathrm{T}}\right] = \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right]\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}
\end{aligned} \tag{53}$$

and recalling from 46 $\Pr(\mathbf{y}) \sim \mathcal{N}(\beta^{\mathrm{T}}\mathbf{X}, \sigma^2\mathbf{I})$

$$\begin{aligned}
&\mathrm{Cov}\left[\mathbf{y}\right] = \sigma^2\mathbf{I} = \\
&= \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] - \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\right]\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\right]
\end{aligned} \tag{54}$$

Rearranging

$$\begin{aligned}
&\Rightarrow \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \sigma^2\mathbf{I} + \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}\right]\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\right] = \\
&= \sigma^2\mathbf{I} + \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{X}\beta\right]\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right] = \\
&= \sigma^2\mathbf{I} + \mathbf{X}\beta\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}
\end{aligned} \tag{55}$$

Substituting:

$$\begin{aligned}
&\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\beta}\hat{\beta}^{\mathrm{T}}\right] = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\left(\sigma^2\mathbf{I} + \mathbf{X}\beta\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\right)\mathrm{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} = \\
&= \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}
\end{aligned} \tag{56}$$

and finally variance-covariance matrix of the least square parameters is

$$\mathrm{Cov}\left[\hat{\beta}\right] = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\sigma^2 \tag{57}$$

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2 \tag{58}$$

3.4 Noise variance estimation

We can find an estimation of the noise variance from the maximum likelihood model using the same procedure used to find the parameters, i.e., taking the derivative and equating it to 0:

$$\begin{aligned}
&\frac{\partial \log \mathrm{L}}{\partial \sigma} = \sum_{\mathrm{i}=1}^{\mathrm{N}} -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(\mathrm{y_i} - \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\beta\right)^2 = 0 \Rightarrow \frac{\mathrm{N}}{\sigma} + \frac{1}{\sigma^3}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y_i} - \mathbf{x}^{\mathrm{T}}\beta\right)^2 = 0 \\
&\Rightarrow \hat{\sigma^2} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y_i} - \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\beta\right)^2
\end{aligned} \tag{59}$$

This can be re-expressed as

$$\begin{aligned}
&\Rightarrow \hat{\sigma^2} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y_i} - \mathbf{x}^{\mathrm{T}}\hat{\beta}\right)^2 = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(\mathrm{y_i} - \mathbf{x}_{\mathrm{i}}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)^2 = \\
&= \frac{1}{\mathrm{N}}\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right) = \\
&= \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{X}\cancel{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \\
&= \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\beta}\right)
\end{aligned} \tag{60}$$

Taking the expectation w.r.t. $\Pr(\mathbf{y}|\mathbf{X}, \beta, \sigma^2)$:

$$\begin{aligned}
&\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\sigma}^2\right] = \frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\beta}\right] = \\
&= \frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{y}\right] - \frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right]
\end{aligned} \tag{61}$$

Suppose $\mathbf{t} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, then $\mathbf{E}_{\mathrm{p}(\mathbf{t})}\left(\mathbf{t^T A t}\right) = \mathbf{Tr}(\mathbf{A\Sigma}) + \mu^{\mathbf{T}}\mathbf{A}\mu$ with $\mu = \mathbf{X}\beta$

$$
\begin{aligned}
&\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\sigma}^2\right] = \frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{y}\right] + \\
&-\frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right] = \\
&= \frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{I}_{\mathrm{N}}\mathbf{y}\right] + \\
&-\frac{1}{\mathrm{N}}\mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\mathbf{y}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\right] = \\
&= \frac{1}{\mathrm{N}}\left(\mathbf{Tr}\left(\sigma^2\mathbf{I}_{\mathrm{N}}\right) + \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta\right) + \\
&-\frac{1}{\mathrm{N}}\left(\mathbf{Tr}\left[\sigma^2\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] + \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta\right) = \\
&= \frac{1}{\mathrm{N}}\left(\mathrm{N}\sigma^2 + \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta\right) - \frac{1}{\mathrm{N}}\left(\sigma^2\mathbf{Tr}\left[\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] + \beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta\right) = \\
&= \sigma^2 + \cancel{\frac{1}{\mathrm{N}}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta} - \frac{\sigma^2}{\mathrm{N}}\mathbf{Tr}\left[\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right] \cancel{-\frac{1}{\mathrm{N}}\beta^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathrm{X}\beta} = \\
&= \sigma^2 - \frac{\sigma^2}{\mathrm{N}}\mathbf{Tr}\left[\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\right] = \sigma^2 - \frac{\sigma^2}{\mathrm{N}}\mathbf{Tr}\left[\mathbf{I}_{\mathrm{p}}\right] \\
&\Rightarrow \mathbf{E}_{\Pr(\mathbf{y}|\mathbf{X},\beta,\sigma^2)}\left[\hat{\sigma}^2\right] = \sigma^2\left(1 - \frac{\mathrm{p}}{\mathrm{N}}\right)
\end{aligned}
\tag{62}
$$

where lastly we have used the product property of the trace (see 1.1).

Normally $\mathrm{p} < \mathrm{N}$ hence the estimate of the variance is smaller than the true variance, so this estimator is **biased**. The estimate gets closer to the real value when p/N is small, i.e., assuming p is fixed, increasing the samples used.

This result might be strange. First of all notice from 59 that the closer the model gets to the data, the smaller $\hat{\sigma}^2$. By definition the parameter estimates are the ones that minimise the noise and hence $\hat{\sigma}^2$. As a consequence, when using the true parameters we would get equal or higher variance.

To estimate the true variance we can use the following formula:

$$\hat{\sigma}^2 = \frac{1}{\mathrm{N-p-1}} \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2 \tag{63}$$

$\mathrm{N-p-1}$ makes this estimation unbiased i.e., $\frac{1}{\mathrm{N-p-1}}\mathbf{E}[\hat{\sigma}^2] = \sigma^2$.

3.5 Interpretation of covariance

Consider a covariance matrix of size 2×2 for a two parameter model (i.e., a line on the plane) and suppose the first diagonal element, corresponding to the variable $\hat{\beta}_0$ is much bigger than the second one corresponding to $\hat{\beta}_1$. This means that we can change $\hat{\beta}_0$ a little without affecting too much the model. On the contrary if the variance is small, small changes will affect significantly the model. Sometimes this happens when one variable has a much higher absolute value.

If the values on the off-diagonals are negative, then when increasing one coefficient i.e., $\hat{\beta}_0$, the other must be decreased to have the line to pass as close as possible to all points. For example in 2D, increasing $\hat{\beta}_0$ reduces the coefficient value: if it is positive, the line will be "more horizontal", if negative it becomes steeper, "more vertical".

3.6 Z-score

Let us assume data were really generated by a linear model but were corrupted by Gaussian noise with 0 mean and variance σ^2:

$$\mathrm{Y} = \beta_0 + \sum_{\mathrm{i}}^{\mathrm{p}} \beta_\mathrm{i} \mathrm{X_i} + \epsilon \tag{64}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The estimated parameters will still be a normal distribution:

$$\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\sigma^2) \tag{65}$$

Definition 3.1. Z-score A Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. Z-scores are measures of an observation's variability

$$\mathrm{z_j} = \frac{\mathrm{x} - \mu}{\sigma} \tag{66}$$

In machine learning the z-value is the regression coefficient divided by its standard error. It is also sometimes called the z-statistic. If the z-value is too big in magnitude (i.e., either too positive or too negative), it indicates that the corresponding true regression coefficient is not 0 and the corresponding X-variable matters. A good rule of thumb is to use a cut-off value of 2 which approximately corresponds to a two-sided hypothesis test with a significance level of $\alpha = 0.05$.

Z-values are computed as the test statistic for the hypothesis test that the true corresponding regression coefficient β is 0. In hypothesis testing, we assume the null hypothesis is true, and then see if the data provide evidence against it. So in this case, we assume β is 0. That is, we assume the expectation of the fitted regression coefficient $\hat{\beta}$ is 0:

$$\mathrm{z_j} = \frac{\hat{\beta}_\mathrm{j}}{\hat{\sigma}\sqrt{\mathrm{v_j}}} \tag{67}$$

where the denominator is the variance of the parameter (from 65) with $\mathrm{v_j}$ being the diagonal element of $(\mathrm{X^T X})^{-1}$.

Theorem 1 (**The Gauss-Markov theorem**)**.** Among all linear unbiased estimators, the least square estimates of the parameters are the ones having smallest variance and consequently from 37 is the one with the smallest mean squared error (the bias-squared term for unbiased estimator is by definition 0).

Proof. Let $\hat{\beta} = \mathrm{Cy}$ be another linear estimator of β with $\mathrm{C} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}$

$$\mathbf{E}\left[\hat{\beta}\right] = \mathbf{E}\left[\mathrm{Cy}\right] = \mathbf{E}\left[\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}\right)(\mathrm{X}\beta + \epsilon)\right] = \tag{68}$$

$$= \left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}\beta + \mathrm{DX}\beta\right) + \cancel{\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}\right)\mathbf{E}[\epsilon]} = \tag{69}$$

$$= (\beta + \mathrm{DX}\beta) = (\mathrm{I} + \mathrm{DX})\,\beta \tag{70}$$

$$\tag{71}$$

where $\mathbf{E}[\epsilon] = 0$.

To be an unbiased estimator $\mathrm{DX} = 0$, then

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}(\mathrm{Cy}) = \mathrm{C}\mathrm{Var}(\mathrm{y})\mathrm{C}^{\mathrm{T}} = \sigma^2\mathrm{C}\mathrm{C}^{\mathrm{T}} = \\
&= \sigma^2\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}\right)\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}\right)^{\mathrm{T}} = \\
&= \sigma^2\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}} + \mathrm{D}\right)\left(\mathrm{X}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \mathrm{D}^{\mathrm{T}}\right) = \\
&= \sigma^2(\cancel{\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{D}^{\mathrm{T}} + \\
&\quad + \mathrm{DX}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \mathrm{D}\mathrm{D}^{\mathrm{T}}) = \\
&= \sigma^2(\cancel{\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{X}}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}(\mathrm{DX})^{\mathrm{T}} + \\
&\quad + \mathrm{DX}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \mathrm{D}\mathrm{D}^{\mathrm{T}})
\end{aligned}$$

$$\begin{aligned}
&\mathrm{DX} = 0 \\
\Rightarrow\ &\mathrm{Var}(\tilde{\beta}) = \sigma^2\left(\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1} + \cancel{\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}(\mathrm{DX})^{\mathrm{T}}} + \cancel{\mathrm{DX}\left(\mathrm{X}^{\mathrm{T}}\mathrm{X}\right)^{-1}} + \mathrm{D}\mathrm{D}^{\mathrm{T}}\right) \\
&\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(\hat{\beta}) + \sigma^2\mathrm{D}\mathrm{D}^{\mathrm{T}}
\end{aligned} \tag{72}$$

□

3.7 Orthogonalization

Normally inputs are not perpendicular but can be orthogonalized. The goal is to define a new orthogonal basis for the data. The procedure, named

Grand-Schmidt procedure is the following:

initialize $z_0 = x_0 = \mathbf{1}$ where $\mathbf{1}$ is a vector of all ones;
For $j = 1, \cdots, p$ regress x_j on $z_0, \cdots, z_{p-1}$ to produce the coefficients $\hat{\gamma}_{lj} = \frac{\langle z_l, x_j \rangle}{\langle z_l, z_l \rangle}$ for $l = 0, \cdots, j-1$;
Calculate the residual vectors as $z_j = x_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj} z_k$;
Regress $\mathbf{y}$ on the residual z_p to get the estimate $\hat{\beta}_p = \frac{\langle z_p, y \rangle}{\langle z_p, z_p \rangle}$

When inputs are correlated, the residual will be close to zero generating instabilities in the coefficients $\hat{\beta}_j$ and the z-score will be small.

The algorithm in matrix form is

$$\mathbf{X} = \mathbf{Z\Gamma}$$

where z_j are the column vectors of Z and Γ is upper triangular. Introducing the diagonal matrix D with j-th diagonal element $D_{jj} = ||z_j||$ we get:

$$\mathbf{X} = \mathbf{ZD^{-1}D\Gamma} = \mathbf{QR} \tag{73}$$

where Q is an orthogonal, $Q^T Q = I$, $N \times (p+1)$ matrix and R is a $(p+1)(p+1)$ upper triangular matrix.

The least square solution becomes

$$\hat{\beta} = R^{-1} Q^T y \hat{y} = Q Q^T y \tag{74}$$

3.8 Multivariate output

We can rewrite the equation in matrix form:

$$Y = XB + E \tag{75}$$

where Y is a $N \times K$ matrix, X is $N \times (p+1)$, B is $(p+1) \times K$, E has the same dimensions of Y. The root squared error is

$$\begin{aligned} RSS(B) &= \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f(x_i))^2 = \mathrm{tr}\left[(Y - XB)^T (Y - XB) \right] \\ \Rightarrow B &= \left(X^T X \right)^{-1} X^T Y \end{aligned} \tag{76}$$

Multiple outputs do not affect one another's least square estimates.

3.9 Subset selection

Least square estimates sometimes show low bias but high variance resulting in a non-satisfactory prediction accuracy. This can be improved by setting some coefficients to 0 to sacrifice a little of bias to reduce significantly the variance.

Other times it is useful to reduce the input dimensionality for easiness of interpretation or for computation purposes.

3.9.1 Best subset selection

This algorithm finds for each $k \in 0, 1, 2, \cdots p$, the subset of size k that gives the smallest residual sum of squares. Note that if a variable is in the best subset of size m, it might not be in the subsets of larger size (and of course neither in the smallest ones).

3.9.2 Forward stepwise selection

Searching for all the subsets is too computationally intensive (and infeasible for $p > 40$). The *forward stepwise* algorithm starts with the intercept and then sequentially adds to the model the predictor that most improves the fit. QR decomposition can be exploited to choose the next candidate.

This algorithm is a greedy sub-optimal algorithm. Statistically it will have lower variance but higher bias.

3.9.3 Backward stepwise selection

Backward stepwise selection starts with the full model and sequentially removes the predictor having least impact on the model, i.e., having the smallest $Z-$score.

Note: this algorithm can only be applied when $N > p$ while *forward stepwise* can always be used.

3.9.4 Implementations

[From ESLII pg. 60] Some software packages implement hybrid stepwise-selection strategies that consider both forward and backward moves at each step, and select the "best" of the two. For example in the R package the step function uses the AIC criterion for weighing the choices, which takes proper

account of the number of parameters fit; at each step an add or drop will be performed that minimizes the AIC score.

Other more traditional packages base the selection on F-statistics, adding "significant" terms, and dropping "non-significant" terms. These are out of fashion.

3.9.5 Forward stagewise regression

As forward stepwise, it starts with the intercept. At each step the algorithm identifies the variable most correlated with the residual and computes the linear regression coefficient of the residual on this chosen variable and then adds it to the current coefficient for that variable. This is continued till none of the variables have correlation with the residual.

At each step only one coefficient is updated by a small step so that the number of steps is bigger than p. This slow-fitting pays in high dimensions.

Note: forward stagewise is very competitive in

3.10 Shrinkage methods

Subset selection methods either keep or remove a predictor. It has higher variance. Shrinkage or regularization methods are more continuous. They force the model to keep the weights as small as possible.

3.10.1 Ridge regression

Ridge regression shrinks the coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squared errors

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{\text{i}=1}^{\text{N}} \left(\text{y}_\text{i} - \beta_0 - \sum_{\text{j}=1}^{\text{p}} \text{x}_{\text{ij}} \beta_\text{j} \right)^2 + \lambda \sum_{\text{j}=1}^{\text{p}} \beta_\text{j}^2 \right\} \tag{77}$$

where λ is a parameter controlling the amount of shrinkage: the bigger the value the greater the amount of shrinkage. This concept is also used in

the Neural Networks. Another way to express 77 is the following:

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \arg\min_{\beta} \left\{ \sum_{\text{i}=1}^{\text{N}} \left(\text{y}_\text{i} - \beta_0 - \sum_{\text{j}=1}^{\text{p}} \text{x}_{\text{ij}}\beta_\text{j} \right)^2 \right\} \\ \text{subject to} &\sum_{\text{j}=1}^{\text{p}} \beta_\text{j}^2 \leq \text{t}\end{aligned} \tag{78}$$

where there is a one-to-one correspondence between λ and t. Note that the penalization term does not consider β_0 otherwise the procedure will depend on the chosen origin for Y.

This algorithm solves the problem of high variance in case of correlated inputs when big coefficients of correlated variables can be cancelled out. With a constraints on the coefficients this problem is alleviated.

The coefficients are not preserved when the input is scaled. Generally inputs are standardized before applying the algorithm.

$$\text{RSS}(\lambda, \beta) = (\text{y} - \text{X}\beta)^\text{T} (\text{y} - \text{X}\beta) + \lambda\beta^\text{T}\beta \tag{79}$$

$$\hat{\beta}^{\text{ridge}} = \left(\text{X}^\text{T}\text{X} + \lambda\text{I} \right)^{-1} \text{X}^\text{T}\text{y} \tag{80}$$

Now even if X is not full rank, the problem is non singular (the inverse exists). In case of orthonormal inputs, the ridge coefficients are the same of least square but scaled: $\hat{\beta}^{\text{ridge}} = \frac{\hat{\beta}}{1+\lambda}$.

The parameter λ can also be derived assuming a prior distribution $\text{y}_\text{i} \sim \mathcal{N}\left(\beta_0 + \text{x}_\text{i}^\text{T}\beta, \sigma^2\right)$ and the parameters β_j are distributed as $\mathcal{N}\left(0, \tau^2\right)$. Then from 77 $\lambda = \frac{\sigma^2}{\tau^2}$.

Applying the SVD decomposition of the matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathbf{T}$ where X is N × p, U is N × pand V is p × p, the latter two both orthogonal with the columns of U spanning the column space of X and the columns of V spanning the row space of X. D is a p × p diagonal matrix with the elements $\text{d}_1 \geq \text{d}_2 \geq \cdots \geq \text{d}_\text{p} \geq 0$ called singular value decomposition of X. If any $\text{d}_\text{j} = 0$ then X is singular.

The least squares equation can be rewritten as

$$\mathbf{X}\hat{\beta}^{\text{ls}} = \mathbf{X}\left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathbf{T}\mathbf{y} = \mathbf{U}\mathbf{U}^\mathbf{T}\text{y} \tag{81}$$

In case of the ridge regression, the coefficients are

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}\left(\mathbf{X^T X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X^T y} = \mathbf{U U^T y} = \\ &= \sum_{\text{j–1}}^{\text{p}} \mathbf{u}_{\text{j}} \frac{\text{d}_{\text{j}}^2}{\text{d}_{\text{j}}^2 + \lambda} \mathbf{u_j}^{\text{T}} \text{y}\end{aligned} \tag{82}$$

where u_{j} are the column vectors. So ridge regression first computes the coordinates of y with respect to the orthonormal basis $\mathbf{U}$, it then shrinks those coordinates since $\lambda \geq 0$. A greater amount of shrinkage is applied to the coordinates of basis vector with smaller d_{j}, corresponding to elements with small variance.

3.10.2 Lasso regression

The lasso regression is similar to ridge regression but it uses a L_1 penalization instead of L_2.

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{\text{i=1}}^{\text{N}} \left(\text{y}_{\text{i}} - \beta_0 - \sum_{\text{j=1}}^{\text{p}} \text{x}_{\text{ij}}\beta_{\text{j}} \right)^2 + \lambda \sum_{\text{j=1}}^{\text{p}} |\beta_{\text{j}}| \right\} \tag{83}$$

or equivalently

$$\begin{aligned}&\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{\text{i=1}}^{\text{N}} \left(\text{y}_{\text{i}} - \beta_0 - \sum_{\text{j=1}}^{\text{p}} \text{x}_{\text{ij}}\beta_{\text{j}} \right)^2 \right\} \\ &\text{subject to} \sum_{\text{j=1}}^{\text{p}} |\beta_{\text{j}}| \leq \text{t}\end{aligned} \tag{84}$$

If $\text{t} > \text{t}_0 = \sum_1^{\text{p}} |\beta_{\text{j}}|$ where β_{j} are the least square coefficients, then the lasso coefficients are the same of the least squares ones. If $\text{t} = \text{t}_0/2$ then the least square coefficients are shrunk by 50% on average. Making t sufficiently big will cause some of the coefficients to be exactly 0.

In 1, the blue areas show the constraints for the estimates of ridge and lasso regressions, while the ellipses are contours of the residual sum of squares, centered at β^{ls}, in case of only two coefficients. For the lasso one of the coefficients is 0 when it hits one of the corner. In higher dimensions the

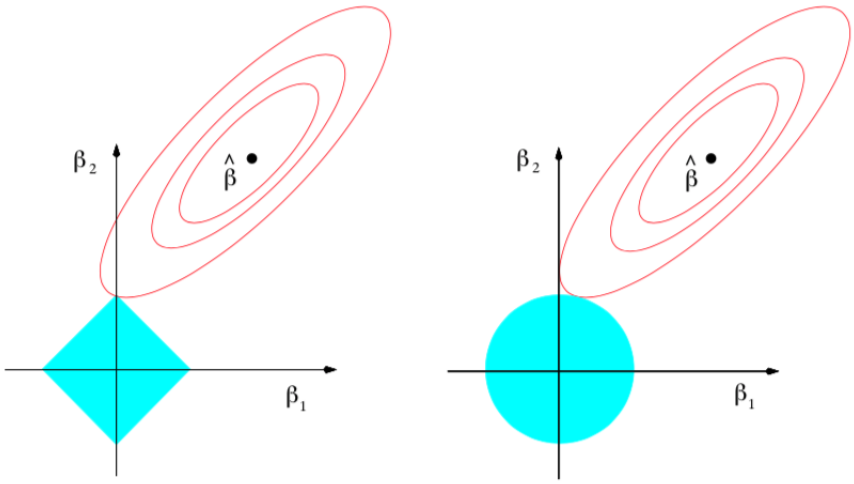


Figure 1: Constraints of ridge and lasso regression.

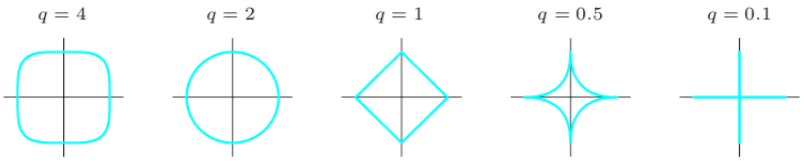


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Figure 2: Shapes of different penalization factors.

figure becomes a rhomboid with many corners (and faces) so it becomes easy to hit a corner.

Also other penalization factors can be chosen, 2

3.10.3 Least angle regression

It is similar to forward stepwise. At first it identifies the variable most correlated with the response but instead of fitting it completely, it moves the variable toward its least squares value. As soon as another value gets correlated to the residual as the previous variable, the process is paused. The second variable joins the active set and their coefficients are moved together in a way that keeps their correlations tied and decreasing. The process is continued until all the variables are in the model. After $\min(N-1, p)$ steps

standardize the predictors to have 0 mean and unit norm;
$\mathcal{A}_k \leftarrow 0$;
$r \leftarrow y - \bar{y}$, $\beta_i = 0$ for $i \neq 0$;
while $|\mathcal{A}_k| < p$ **do**
  find the predictor most correlated to r;
  insert the predictor in the active set $\mathcal{A}_k$;
  move the coefficients of the predictors in the active set to the direction defined by their joint least squares coefficient of the current residual, i.e.,

$$\delta_k = \left(X_{\mathcal{A}_k}^T X_{\mathcal{A}_k}\right)^{-1} X_{\mathcal{A}_k}^T r_k \tag{85}$$

  (where $\mathcal{A}_k$ is the current active set of variables) until some other competitor x_l has as much correlation with the current residual;
  $r_k \leftarrow y - X_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$
end

we we arrive at the full least squares solution. The coefficient profile evolves as

$$\beta_{\mathcal{A}_k} = \beta_{\mathcal{A}_k} + \alpha\delta_k \tag{86}$$

If to the LAR algorithm we add the following rule i.e.,

If a non-zero coefficient hits 0*, drop its variable from the active set of variables and recompute the current joint least squares direction.*

we get the same coefficient path of the lasso and this is called LAR(lasso). So this become an efficient solution to compute the Lasso problem, especially with $N >> p$ since Lasso can take more than p steps while LAR require p steps and it is efficient since it requires the same complexity as that of a single least squares fit using the p predictors.

3.11 Derived input directions methods

When a large number of inputs is present, often there is a high correlation among them. In this case is convenient to regress on a new set inputs obtained form a linear combination of the original input.

3.11.1 Principal component analysis

First input must be standardized since this analysis depends on the scaling. The principal components are defined as

$$z_i = Xv_i \tag{87}$$

where v_i are the column vectors of V from the SVD decomposition of $X = UDV^T$ (recall that z_m are orthogonal). The algorithm then regress X on $z_1, \cdots, z_M$ for $M \leq p$ and we have:

$$\hat{y}^{pcr}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m \tag{88}$$

where $\hat{\theta}_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$. Since z_m are linear combination of the original predictors x_j, we can express the solution as

$$\hat{\beta}^{pcr}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m \tag{89}$$

With $M = p$ we get the usual least squares. Principal component analysis discards the $p - M$ smallest eigenvalue components.

The value M is suggested by cross-validation.

3.11.2 Partial least squares

This technique use a set of linear combinations of y in addition to X for the construction. It is not scale invariant so each x_j must be standardized.

Since PLS use y to construct its directions, its solution path is not linear in y. It seeks direction with high variance and high correlation with the response while PCR only with high variance.

3.12 Multioutput shrinkage and selection

To apply selection and shrinkage methods in the multiple output case, one could apply a univariate technique individually to each outcome or si- multaneously to all outcomes, i.e., different λ in Ridge or Lasso can be used for each output or the same value can be adopted.

standardize x_j to 0 mean and 1 variance;
$\hat{y}^{(0)} \leftarrow \bar{y}\mathbf{1}$;
$x_j^{(0)} \leftarrow x_j$;
for $m = 1, \cdots, p$ **do**
  $z_m = \sum_{j=1}^{p} \hat{\phi}_{mj} x_j^{(m-1)}$ where $\hat{\phi}_{mj} = \langle x_j^{(m-1)}, y \rangle$;
  $\hat{\theta}_m = \frac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}$;
  $\hat{y}^m = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$;
  orthogonalize each x_j^{m-1} w.r.t.
    $z_m : x_j^{(m)} = x_j^{(m-1)} - \frac{\langle z_m, x_j^{(m-1)} \rangle}{\langle z_m, z_m \rangle} z_m$, $j = 1, \cdots, p$;
end
Output the sequence of fitted vectors $\{\hat{y}^m\}_1^p$. Since z_{l1}^m are linear in x_j, so is $\hat{y}^{(m)} = X\hat{\beta}^{pls}(m)$. These coefficients can be recovered from the sequence of PLS transformations.

3.13 Other derived algorithms

3.13.1 Incremental forward stagewise

$r \leftarrow y$;
$\beta_i = 0$ for $i \neq 0$;
find the (standardized) predictor x_j most correlated with the residual ;
$\beta_j \leftarrow \beta_j + \delta_j$ where $\delta_j = \epsilon \text{sign}\left[\langle x_j, r \rangle\right]$ and $\epsilon > 0$ small;
$r \leftarrow r - \delta_j x_j$;
Repeat the steps many times until the residuals are uncorrelated with the predictors.

3.13.2 The Dantzig selector

...

3.13.3 The Grouped Lasso

...

4 Bayesian inference

4.1 Introduction to Bayes' theorem

Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Given a hypothesis H and evidence E, Bayes' theorem states that the relationship between the probability of the hypothesis Pr(H), before getting the evidence, and the probability of the hypothesis after getting the evidence Pr(H|E) is

$$\Pr(\mathrm{H}|\mathrm{E}) = \frac{\Pr(\mathrm{E}|\mathrm{H})\Pr(\mathrm{H})}{\Pr(\mathrm{E})} \tag{90}$$

Often there are competing hypothesis and the task is to determine which is the most probable.

This formula relates the probability of the hypothesis before getting the evidence Pr(H), to the probability of the hypothesis after getting the evidence, Pr(H|E): this term is generally what we want to know. For this reason, Pr(H) is called the **prior probability**, while Pr(H|E) is called the **posterior probability**. The factor that relates the two, $\frac{\Pr(\mathrm{E}|\mathrm{H})}{\Pr(\mathrm{E})}$, is called the **likelihood ratio** and Pr(E|H) is called **likelihood** which indicates the compatibility of the evidence with the given hypothesis. P(H|E) and Pr(E|H) are called conditional probabilities. A conditional probability is an expression of how probable one event is given that some other event occurred (a fixed value) and Bayes' theorem centers on relating different conditional probabilities.

Pr(E) is sometimes called **marginal likelihood** or model evidence this factor is the same for all the hypotheses being considered.

4.2 Bayes inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference assumes the data were generated by a model with unknown parameters. From this, it tries to come up with beliefs about the likely "true values" of the parameters of the model. For this reason when writing down the Bayes theorem in these cases, formally also the conditional on the choice of the model should be written down:

$$\Pr(\theta|\mathbf{X}, \mathrm{M}) = \frac{\Pr(\mathbf{X}|\theta, \mathrm{M})\Pr(\theta|\mathrm{M})}{\Pr(\mathbf{X}|\mathrm{M})} \tag{91}$$

Here the hypothesis described in the previous paragraph is represented by a set of values for the parameters.

Once the model is stipulated, $\Pr(\mathrm{X}|\theta)$ can be evaluated for any given set of parameters: in this sense the likelihood is fixed once the model is fixed. $\Pr(\mathbf{X}|\mathrm{M})$ can be rexpressed as $\Pr(\mathbf{X}|\mathrm{M}) = \int \Pr(\mathbf{X}|\theta, \mathrm{M})\Pr(\theta|\mathrm{M})\mathrm{d}\theta$. In this way it can be re-thought as a normalization factor of the sets of parameters since it is a summation over the all parameter space. However note that it **does depend** on the choice of the model. It can be seen also as asking the question *how much is it probable to see the data we have seen given that the model* M*, without any claim about its parameters, generated those data*?

4.3 Types of estimation

It is the moment to clarify different types of estimation. We have already seen the Ordinary Least Square (OLS) estimation, and other types of estimation based on the definition of an error function.

We have also seen the **Maximum Likelihood Estimation (MLE)** approach, which maximizes the likelihood of the Bayes expression and how it is equivalent to OLS estimation in case of linear regression. (To be precise almost always the log-likelihood is maximized, first of all because of analytical convenience since many times we deal with exponentials coming from Gaussian distribution. Secondly for numerical precision: we are dealing with probabilities, numbers between 0 and 1 and since the range is quite small, underflow might be a problem.) The logarithmic instead extends the range by mapping numbers close to 0 to ∞, resulting in a better precision.

The **Maximum A-Priori (MAP)** estimation maximizes the numerator of the Bayes expression, i.e., the likelihood times the prior, which means the likelihood is weighted by weights coming from the prior. When using a uniform distribution for the prior, MAP turns into MLE since we are assigning equal weights for each possible value. For example suppose we can assign six possible values to β and assume $\mathrm{P}(\beta_\mathrm{i}) = 1/6$:

$$\begin{aligned}\beta_{\mathrm{MAP}} &= \arg\max_{\beta} \sum_{\mathrm{i}} \Pr(\mathbf{x}_\mathrm{i}|\beta) + \log \mathrm{P}(\beta) = \\ &= \arg\max_{\beta} \sum_{\mathrm{i}} \Pr(\mathbf{x}_\mathrm{i}|\beta) + \mathrm{const} = \arg\max_{\beta} \sum_{\mathrm{i}} \Pr(\mathbf{x}_\mathrm{i}|\beta) = \beta_{\mathrm{MLE}}\end{aligned} \tag{92}$$

When using a different prior, i.e., the simplification does not hold anymore.

4.4 Conjugate distributions

Definition 4.1. Conjugate distributions In Bayesian probability theory, if the posterior distribution $\Pr(\theta|\mathbf{X}, \mathrm{M})$ is in the same probability space as the prior probability distribution $\Pr(\theta|\mathrm{M})$, then the prior and posterior are called **conjugate distributions**, and the prior is called **conjugate prior for the likelihood function**.

As example consider the Gaussian distribution: **the Gaussian family is conjugate to itself (or self-conjugate) with respect to a Gaussian likelihood function**. If the likelihood is Gaussian, choosing Gaussian prior will ensure that also the posterior distribution is a Gaussian (see 1.6). This means that the Gaussian distribution is a conjugate prior for the likelihood that is also Gaussian.

Consider the general problem of inferring a (continuous) distribution for a parameter θ given some datum or data x. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function $\mathrm{p}(\mathrm{x}|\theta, \mathrm{M})$ and the prior $\mathrm{p}(\theta|\mathrm{M})$. Let the likelihood function be considered fixed; the likelihood function is usually well-determined from a statement of the data-generating process. It is clear that different choices of the prior distribution $\mathrm{p}(\theta|\mathrm{M})$ may make the integral more or less difficult to calculate, and the product $\mathrm{p}(\mathrm{x}|\theta, \mathrm{M}) \times \mathrm{p}(\theta|\mathrm{M})$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is a conjugate

prior. A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise numerical integration may be necessary. Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution. All members of the exponential family have conjugate priors.

In case of **classification** we are given a training set with input and output data. Once we have stipulated a model that could have generated output data, we want to find the best parameters of the model such that when those inputs are fed to the model we get those outputs. So the question becomes *what is the best set of parameters* θ *for the model* M *that could have generated the output* $\mathbf{y}$ *when the model has been fed with input data* $\mathbf{X}$*?*

In formula:

$$\Pr(\theta|\mathbf{y}, \mathbf{X}, \mathrm{M}) = \frac{\Pr(\mathbf{y}|\theta, \mathbf{X}, \mathrm{M})\Pr(\theta|\mathbf{X}, \mathrm{M})}{\Pr(\mathbf{y}|\mathbf{X}, \mathrm{M})} \tag{93}$$

4.4.1 Dataset likelihood

4.10.1 To be precise the likelihood is the likelihood of an entire dataset, since we are interested in all $\mathbf{y}$ and not in a single value y. $\Pr(\mathbf{y}|\theta, \mathbf{X}, \mathrm{M})$ is then a joint density over all the responses in our dataset: $\Pr(\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_\mathrm{n}|\theta, \mathbf{X}, \mathrm{M})$. Evaluating this density at the observed points gives a single likelihood value for the whole dataset. Assuming that the noise at each data point is independent we can factorize as:

$$\Pr(\mathbf{y}|\theta, \mathbf{X}, \mathrm{M}) = \prod_{\mathrm{n}=1}^{\mathrm{N}} \Pr(\mathrm{y}_\mathrm{n}|\mathbf{x}_\mathrm{n}, \theta) \tag{94}$$

We have not say that y_n's are completely independent as otherwise it would not be worth trying to model the data at all. Rather, they are **conditionally independent** given a value θ, i.e., the deterministic model. Basically the model incorporates the dependency. Consider the following example, we have a set of data $(\mathbf{y}, \mathbf{X})$ from which we want to predict the output y_n given a new $\mathbf{x}_\mathrm{n}$. Recalling from conditional probability that $\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{A} \cap \mathrm{B})}{\Pr(\mathrm{B})}$ we have:

$$\mathrm{P}(\mathrm{y}_\mathrm{n}|\mathbf{y}, \mathbf{x}_\mathrm{n}, \mathbf{X}) = \frac{\Pr(\mathrm{y}_\mathrm{n} \cap \mathbf{y}|\mathbf{x}_\mathrm{n}, \mathbf{X})}{\Pr(\mathbf{y}|\mathbf{x}_\mathrm{n}, \mathbf{X})} \tag{95}$$

Note that actually $\mathbf{y}$ does not depend on $\mathbf{x}_\mathrm{n}$: $\Pr(\mathbf{y}|\mathbf{x}_\mathrm{n}, \mathbf{X}) = \Pr(\mathbf{y}|\mathbf{X})$. Using independence:

$$\mathrm{P}(\mathrm{y}_\mathrm{n}|\mathbf{y}, \mathbf{x}_\mathrm{n}, \mathbf{X}) = \frac{\Pr(\mathrm{y}_\mathrm{n} \cap \mathbf{y}|\mathbf{x}_\mathrm{n}, \mathbf{X})}{\Pr(\mathbf{y}|\mathbf{X})} = \frac{\Pr(\mathbf{y}|\mathbf{X})\Pr(\mathrm{y}_\mathrm{n}|\mathbf{x}_\mathrm{n}, \mathbf{X})}{\Pr(\mathbf{y}|\mathbf{X})} = \tag{96}$$

[TO BE CONTINUED WAITING ON ANSWER ON CROSS-VALIDATED STACK EXCHANGE]

5 Linear Classification

For classification problem, in this section we assume the classification boundaries are linear, i.e., in the input hyperspace the points belonging to different classes can be separated by hyperplanes.

5.1 Linear regression of an Indicator matrix

Suppose we have K classes. For a single output we build a vector $\mathbf{y} = (\mathrm{y}_1, \cdots, \mathrm{y}_\mathrm{k})$ where $\mathrm{y}_\mathrm{k} = 1$ if the class it belongs is the k class. As output we will have the matrix $\mathbf{Y}$ of 0 and 1 with each row having a single 1. We fit a linear regression model to each of the columns of $\mathbf{Y}$ simultaneously:

$$\hat{\mathbf{Y}} = \mathbf{X}\left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathrm{T}\mathbf{Y}$$

We get a coefficient vector for each response column y_k, and hence a $(\mathrm{p}+1)\times\mathrm{K}$ matrix $\hat{\mathbf{B}} = \left(\mathbf{X}^\mathrm{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathrm{T}\mathbf{Y}$,where $\mathbf{X}$ will have $\mathrm{p}+1$ columns with a leading column of 1 for the intercept.

Suppose we are given a new input x. Then the classification problem becomes:

compute the output $\hat{\mathrm{f}}(\mathrm{x})^\mathrm{T} = (1, \mathrm{x}^\mathrm{T})\hat{\mathbf{B}}$ *which is a* k *vector identify the largest component* $\hat{\mathrm{G}}(\mathrm{x}) = \arg\max_{\mathrm{k}\in\mathcal{G}} \hat{\mathrm{f}}_\mathrm{k}(\mathrm{x})$.

With this approach we are basically estimating a conditional expectation, i.e., given the inputs $\mathbf{x}$ what is the probability the output is of class k? Mathematically $\mathbf{E}\left(\mathrm{y}_\mathrm{k}|\mathrm{X} = \mathrm{x}\right) = \Pr\left(\mathrm{G} = \mathrm{k}|\mathrm{X} = \mathrm{x}\right)$ since $\mathrm{y}_\mathrm{k} = 1$.

Although the linear model guarantees $\sum_{\mathrm{k}\in\mathcal{G}} \hat{\mathrm{f}}_\mathrm{k} = 1$, as long as there is an intercept in the model, $\hat{\mathrm{f}}_\mathrm{k}(\mathrm{x})$ can be negative or bigger than one, especially

when making predictions outside the hull of training data. Although this fact, this approach still works in many cases.

An important limitation is when $K \geq 3$. Even if the classes can still be separated by more than one linear boundaries, linear regression cannot find linear boundaries (because they are more than one?).

Quadratic regression might solve the problem, but a general rule is that if $k \geq 3$ classes are lined up (their centroids are in the same line), a polynomial term with degree up to $k - 1$ is needed to solve the problem, and since the direction is arbitrary, cross-products terms might be needed too.

5.2 Linear Discriminant analysis

For optimal classification we have to know the class posteriors $\Pr(G|X)$. Let π_k be the prior probability of class k with $\sum_k^K \pi_k = 1$. Suppose $f_k(x)$ is the class conditional density. From the Bayes theorem (?) we get

$$\Pr(G = k|X = x) = \frac{\Pr(X = x|G = k)\Pr(G = k)}{\Pr(X = x)} \tag{97}$$

$\Pr(G = k)$ is the prior probability π_k, $\Pr(X = x|G = k)$ is the class conditional density while the denominator can be rewritten using the **Total Probability Theorem** as

$$\Pr(X = x) = \sum_{l=1}^{K} \Pr(X = x|G = l)\Pr(G = l) = \sum_{l=1}^{K} \Pr(X = x|G = l)\pi_l \tag{98}$$

$$\Rightarrow \Pr(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \tag{99}$$

The goodness of classification mostly rely on $f_k(x)$ and many techniques use models for class densities:

- linear and quadratic discriminant analysis use Gaussian densities;
- mixtures of Gaussians allow for non-linear boundaries;
- Naive Bayes models assume that each class density is a product of marginal densities i.e., inputs are conditionally independent in each class.

Modelling each class density as a multivariate Gaussian we have

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\sum_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)} \tag{100}$$

Linear Discriminant analysis assumes equal covariance matrices for all classes. Taking as comparison between two classes the log-ratio, we have

$$\begin{aligned}
&\log \frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = \log \frac{f_k(x)\pi_k}{f_l(x)\pi_l} = \log \frac{\pi_k}{\pi_l} + \log \frac{f_k(x)}{f_l(x)} = \\
&= \log \frac{\pi_k}{\pi_l} + \log \frac{e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}}{e^{-\frac{1}{2}(x-\mu_l)^T \Sigma^{-1}(x-\mu_l)}} = \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) - \left(-\frac{1}{2}\right)(x-\mu_l)^T \Sigma^{-1}(x-\mu_l) = \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left[(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + (x-\mu_l)^T \Sigma^{-1}(\mu_l - x)\right] = \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left[(2x-\mu_k-\mu_l)^T \Sigma^{-1}(\mu_l - \mu_k)\right] = \\
&= \log \frac{\pi_k}{\pi_l} + \frac{1}{2} x^T \Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) =
\end{aligned} \tag{101}$$

which is linear in x, so all decision boundaries are linear (i.e., they are hyperplanes in p dimensions). If the common covariance matrix is spherical, i.e., $\Sigma = \sigma^2 I$ and the class priors are equal, each boundary that separates two classes is the perpendicular bisector of the segment joining the centroids of the two classes.

The linear discriminant functions of each class are

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T + \log \pi_k \tag{102}$$

We do not know the parameters of the Gaussian distribution and we must

estimate them from the training data:

$$\hat{\pi}_k = \frac{N_k}{N} \tag{103}$$

$$\hat{\mu}_k = \sum_{g_i=k} \frac{x_i}{N_k} \tag{104}$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{(N - K)} \tag{105}$$

Note that LDA does not use Gaussian assumption for the features.

5.2.1 Decision rule

Consider two classes 1 and 2. LDA classifies to class 1 if

$$x^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_1 - \frac{1}{2} \hat{\mu}_1^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_1 + \log \hat{\pi}_1 > x^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_2 - \frac{1}{2} \hat{\mu}_2^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_2 + \log \hat{\pi}_2 \tag{106}$$

to class 2 if $<$ holds. In such case there is a correspondence between LDA and linear regression classification if the two classes are coded with +1 and –1. In this case the coefficient vector from least squares is proportional to the LDA direction. However unless $N_1 = N_2$, the intercepts are different and so are the decision rules.

With more than 2 classes, linear regression is not able to classify correctly while LDA does.

5.3 Quadratic Discriminant analysis

If we do not assume equal covariance, the squared term in x does not cancel out and we get quadratic discriminant functions:

$$\delta_k(x) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^T \mathbf{\Sigma}_k^{-1} (x - \mu_k) + \log \pi_k \tag{107}$$

The decision boundaries between two classes are quadratic functions.

Note: QDA does not differ much from LDA applied the enlarged quadratic polynomial input space but generally QDA is preferred in this case.

The estimates are similar but the covariance matrix must be estimated for each class. When p is large, this means a dramatic increase in the number of parameters, considering we only need the differences $\delta_k(x) - \delta_l(x)$. LDA needs $(K - 1) \times (p + 1)$ parameters, while QDA needs $(K - 1) \times (p\,(p + 3)/2 + 1)$.

5.4 Regularized discriminant analysis

This method shrinks the separate covariances of QDA towards a common covariance as in LDA. In a way it is similar to ridge regression. The regularized covariance matrices have the form:

$$\hat{\mathbf{\Sigma}}_{\mathrm{k}}(\alpha) = \alpha\hat{\mathbf{\Sigma}}_{\mathrm{k}} + (1-\alpha)\hat{\mathbf{\Sigma}} \tag{108}$$

with $\alpha \in [0,1]$, the two extremes being LDA and QDA. α can be chosen on the validation data or by cross-validation.

Similarly we can allow $\hat{\mathbf{\Sigma}}$ to be shrunk toward the scalar covariance:

$$\hat{\mathbf{\Sigma}}(\gamma) = \gamma\hat{\mathbf{\Sigma}} + (1-\gamma)\hat{\sigma}^2\mathrm{I} \tag{109}$$

with $\gamma \in [0,1]$ so we get a more general family of covariances

$$\hat{\mathbf{\Sigma}}_{\mathrm{k}}(\alpha,\gamma) = \alpha\hat{\mathbf{\Sigma}}_{\mathrm{k}} + (1-\alpha)\left(\gamma\hat{\mathbf{\Sigma}} + (1-\gamma)\,\hat{\sigma}^2\mathrm{I}\right) \tag{110}$$

5.5 Computation

Computation of LDA and QDA is simplified by diagonalizing the covariance matrices with the singular value decomposition $\hat{\mathbf{\Sigma}}_{\mathrm{k}} = \mathrm{U_k D_k U_k^T}$. The terms in 100 become

$$\begin{aligned}
&(\mathrm{x}-\mu_{\mathrm{k}})^{\mathrm{T}}\,\mathbf{\Sigma}_{\mathrm{k}}^{-1}\,(\mathrm{x}-\mu_{\mathrm{k}}) = (\mathrm{x}-\mu_{\mathrm{k}})^{\mathrm{T}}\left(\mathrm{U_k D_k U_k^T}\right)^{-1}(\mathrm{x}-\mu_{\mathrm{k}}) = \\
&= (\mathrm{x}-\mu_{\mathrm{k}})^{\mathrm{T}}\,\mathrm{U_k D_k^{-1} U_k^T}\,(\mathrm{x}-\mu_{\mathrm{k}}) = \left[\mathrm{U_k^T}\,(\mathrm{x}-\mu_{\mathrm{k}})\right]^{\mathrm{T}}\mathrm{D_k^{-1}}\left[\mathrm{U_k^T}\,(\mathrm{x}-\mu_{\mathrm{k}})\right]
\end{aligned} \tag{111}$$

$$\log|\hat{\mathbf{\Sigma}}_{\mathrm{k}}| = \sum_{\mathrm{l}}\log \mathrm{d_{kl}} \tag{112}$$

Considering the above steps, LDA classifier can be seen as performing the following steps:

- sphere the data w.r.t. the common covariance estimate: $\mathrm{X}* \leftarrow \mathrm{D}^{-\frac{1}{2}}\mathrm{U}^{\mathrm{T}}\mathrm{X}$. The common covariance estimate fof X$*$ will now be the identity.
- Classify to the closest centroid in the transformed space, modulo the effect of the class prior probability π_{k}(?)

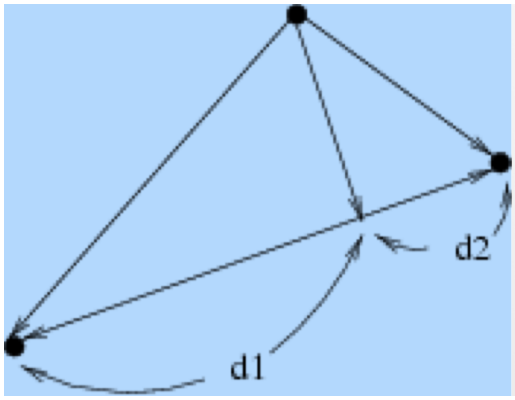


Figure 3: ...

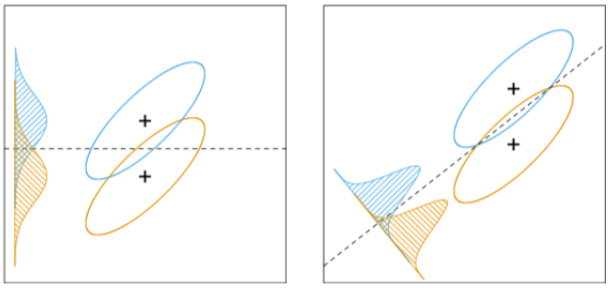
5.6 Regularized-rank linear discriminant analysis

Consider K = 2 with two centroids and the input is 2, i.e., input points are on a plane. Given an input point, for classification purposes what matters is not the distance in the p space of such point from the two centroids but rather the distance from the two centroids of the projection of this point on the line joining them (**??**). So basically instead of using the 2 dimensions, we are calculating the distance in one dimension, a line. If K = 3 then the points are projected onto a plane (2d), of course in this case it is convenient if p > 2.

More generally the K centroids in p-dimensional input, lie in an affine subspace of dimensions $\leq$ K – 1 and if p > is much larger than K this will be a considerable drop in dimension.

If K > 3 we can look for a L < K – 1 dimensional subspace optimal for LDA. Fisher defined *optimal* such that the projected centroids were spread out as much as possible in terms of variance. This problem, finding the principal component subspaces of the centroids, involves the following steps:

- compute the k × p matrix of class centroids M and the common covariance matrix W for within-class covariance;
- compute $M* = MW^{-\frac{1}{2}}$ using the eigen-value decomposition;
- compute B∗, the covariance matrix of M∗ (B for between class covariance) and its eigenvalue decomposition $B^* = V^* D_B V^{*T}$. The columns v_l^* from the first to the last define the coordinates of optimal subspaces.



The lth discriminant variable is given by $Z_l = v_l^T X$ with $v_l = W^{-\frac{1}{2}} v_l^*$. Although the direction joining the centroids separates the means as much as possible(maximizes the between class covariance), there is an overlap between the projected classes due to the nature of covariances. Taking the covariances into account reduce the overlap and that is what we are doing (**??**).

The between-class variance Z is $a^T Ba$ and the within class variance is $a^T Wa$, with $B + W = T$, the total covariance matrix of X. Fisher's problem maximizes the *Rayleigh quotient*:

$$\max_a \frac{a^T Ba}{a^T Wa} \tag{113}$$

This is a generalized eigenvalue problem, with a given by the largest eigenvalue. Similarly one can find the next direction a_2, orthogonal in W to a_1, such that $a_2^T Ba_2 / a_2^T Wa_2$ is maximized; the solution is $a_2 = v_2$, and so on. a_l are the *discriminant coordinates* or *canonical variates*, different from discriminant functions.

The reduced subspaces can be used both for visualization and classification by limiting the distance between centroids to the chosen subspace. However, when doing this, due to the Gaussian classification, a correction factor of $\log \pi_k$ is needed. The misclassification is given by the overlapping area in **??** between the two densities. When both classes have the same priors π_k as in the figure, the optimal cut-point is the midway between projected means, if not the cut-point is moved towards the smaller class to have a better error rate.

For 2 classes one can derive the linear rule using LDA, and then choosing the cut-point to minimize misclassification error.

5.7 Logistic regression

The idea behind logistic regression is to still exploit a linear model $\mathrm{x}^{\mathrm{T}}\beta$ but having its output representing a probability, i.e., constrained between 0 and 1. This part is performed using the sigmoid function

$$\mathrm{p} = \sigma(\mathrm{q}) = \frac{1}{1 + \mathrm{e}^{-\mathrm{q}}} \tag{114}$$

Inverting the terms we get

$$\mathrm{q} = -\log \frac{1-\mathrm{p}}{\mathrm{p}} = \log \frac{\mathrm{p}}{1-\mathrm{p}} \tag{115}$$

115 is called **logit function**. As q increases to ∞, the output of the sigmoid gets closer to 1; instead when it diverges to $-\infty$ we get 0. Suppose we have just 2 classes or equivalently a binary classifier that tells the probability of an event to happen: $\mathrm{Y_n} = 1$ when the event happens and $\mathrm{Y_n} = 0$ when it does not. We can express the probabilities output by our classifier when new input data $\mathrm{x_{new}}$ are observed as:

$$\begin{aligned} \mathrm{P}(\mathrm{G}=1|\mathrm{x_n},\beta) &= \frac{1}{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}} \\ \mathrm{P}(\mathrm{G}=0|\mathrm{x_n},\beta) &= 1 - \frac{1}{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}} = \frac{\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}}{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}} \end{aligned} \tag{116}$$

These equations can be combined in a single equation:

$$\mathrm{P}(\mathrm{G}=\mathrm{g}|\mathrm{x_n},\beta) = \mathrm{P}(\mathrm{G}=1|\mathrm{x_n},\beta)^{\mathrm{g}}\mathrm{P}(\mathrm{G}=0|\mathrm{x_n},\beta)^{\mathrm{g}-1} \tag{117}$$

Taking the log-ratio between the two probabilities we have:

$$\log \frac{\mathrm{P}(\mathrm{Y_n}=0|\mathrm{x_n},\beta)}{\mathrm{P}(\mathrm{Y_n}=1|\mathrm{x_n},\beta)} = \log \frac{\frac{\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}}{\cancel{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}}}}{\frac{1}{\cancel{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}}}} = -\beta^{\mathrm{T}}\mathbf{x}_{\mathrm{new}} \tag{118}$$

So we are using lines (or hyperplanes) to separate the two classes.

5.7.1 Multinomial logistic regression: more than 2 classes

Now suppose we have more than two classes and we still want to separate those classes with linear functions, which means we will have a hyperplane separating two classes. We have K possible outcomes. We can think to run $\mathrm{K}-1$ independent binary logistic regressions with one class chosen as **pivot**, generally the one corresponding to class K, and with the other $\mathrm{K}-1$ classes separately regressed against the pivot:

$$\begin{aligned}
&\log \frac{\Pr(\mathrm{G}=1|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} = \beta_1^{\mathrm{T}}\mathrm{x} \\
&\log \frac{\Pr(\mathrm{G}=2|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} = \beta_2^{\mathrm{T}}\mathrm{x} \\
&\vdots \\
&\log \frac{\Pr(\mathrm{G}=\mathrm{K}-1|\mathrm{X}=\mathrm{x})}{\Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})} = \beta_{(\mathrm{K}-1)}^{\mathrm{T}}\mathrm{x}
\end{aligned} \tag{119}$$

The choice of the pivot class used as denominator is arbitrarily and the estimates are equivalent under different choices.

Summing the probability of each class we must get 1:

$$\begin{aligned}
&\sum_{\mathrm{l}=1}^{\mathrm{K}} \Pr(\mathrm{G}=\mathrm{l}|\mathrm{X}=\mathrm{x}) = 1 \Rightarrow \Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x}) + \sum_{\mathrm{l}=1}^{\mathrm{K}-1} \Pr(\mathrm{G}=\mathrm{l}|\mathrm{X}=\mathrm{x}) = 1 \\
&\Rightarrow \Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x}) + \sum_{\mathrm{l}=1}^{\mathrm{K}-1} \Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x})\, \mathrm{e}^{\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}} = 1 \\
&\Rightarrow \Pr(\mathrm{G}=\mathrm{K}|\mathrm{X}=\mathrm{x}) = \frac{1}{1+\sum_{\mathrm{l}=1}^{\mathrm{K}-1} \mathrm{e}^{\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}}}
\end{aligned} \tag{120}$$

So we can re-express the probabilities as:

$$\Pr(\mathrm{G}=\mathrm{k}|\mathrm{X}=\mathrm{x}) = \frac{\mathrm{e}^{\beta_{\mathrm{k}}^{\mathrm{T}}\mathrm{x}}}{1+\sum_{\mathrm{l}=1}^{\mathrm{K}-1} \mathrm{e}^{\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}}} \tag{121}$$

Softmax function We do not have anymore the sigmoid function, instead we have used another function named **softmax** that as the sigmoid takes

any real value as input and outputs a value between 0 and 1. The difference mostly relies on the denominator used for normalization factor since the sigmoid is a 2D curve while the softmax can have higher dimensionality (for 2D it corresponds to the sigmoid).

5.7.2 Fitting logistic regression

From now on we will consider a logistic regression for just two classes.

As seen in 110, we want to find the best parameters for the chosen model according to some criteria. First let us apply the Bayes theorem:

$$\Pr\left(\beta|\mathbf{y},\mathbf{X},\mathrm{M}\right)=\frac{\Pr\left(\mathbf{y}|\beta,\mathbf{X},\mathrm{M}\right)\Pr(\beta,\mathrm{M})}{\Pr(\mathbf{y}|\mathbf{X},\mathrm{M})} \tag{122}$$

From now on we will not write the conditional on the model for seek of brevity but we know it exists.

This formula tells we want to find the coefficients given some input and output data. Let us analyse each term:

- $\Pr(\beta)$ (**prior distribution**): this is the prior belief about the parameters without seeing the data. As prior, we will use a Gaussian distribution with 0 mean: $\mathcal{N}(0,\sigma^2\mathbf{I})$. In this way the parameters will depend on σ^2: more formally we should write $\Pr(\beta|\sigma)$ but we will skip. Although the choice of the Gaussian distribution with 0 mean is analytically convenient as seen in 4.10, we will not rely on conjugacy.

- $\Pr\left(\mathbf{y}|\beta,\mathbf{X}\right)$ (**likelihood**): we will assume y are conditionally independent (4.10.1):

  $$\Pr(\mathbf{y}|\mathbf{X},\beta)=\prod_{\mathrm{n}=1}^{\mathrm{N}}\Pr(\mathrm{y_n}|\mathbf{x}_\mathrm{n},\beta) \tag{123}$$

  Since in this case we have a binary variable, instead of a Gaussian random variable, suitable for real values distribution, we will consider a binary random variable $\mathrm{Y_n}$ characterised by the probability of the second class:

  $$\Pr(\mathbf{y}|\mathbf{X},\beta)=\prod_{\mathrm{n}=1}^{\mathrm{N}}\Pr(\mathrm{Y_n}=\mathrm{y_n}|\mathbf{x}_\mathrm{n},\beta) \tag{124}$$

- $\Pr(\mathbf{y}|\mathbf{X})$(**marginal likelihood**): It can be expressed as

$$\Pr(\mathbf{y}|\mathbf{X}) = \int \Pr(\mathbf{y}|\beta, \mathbf{X}) \Pr(\beta) \mathrm{d}\beta \tag{125}$$

So the numerator can be calculated and results in a Gaussian function not in standard form (1.6). The denominator, or marginal likelihood, as we have seen can be expressed as $\Pr(\mathbf{y}|\mathbf{X}) = \int \Pr(\mathbf{y}|\beta, \mathbf{X}) \Pr(\beta) \mathrm{d}\beta$. Mathematically this integral has no close solution since (again see 1.6) and the impossibility of the integral of $\mathrm{e}^{-\mathrm{x}^2}$), unless an ad-hoc prior distribution is chosen.

When we cannot directly compute the posterior density (due to the denominator), we have three options:

1. find the single value β that correspond to the highest value of the posterior. This is equivalent to find the value β that maximize the numerator since the denominator is not a function of β but a numerical value.

2. Approximate $\Pr(\beta|\mathbf{X}, \mathbf{y})$ with some other density that we can compute analytically.

3. Sample directly from the posterior $\Pr(\beta|\mathbf{X}, \mathbf{y})$ knowing only the numerator.

The first method is simple and hence popular but it is not very "Bayesian", since we will make predictions of new data based on a single value and not a distribution. The whole Bayesian theory is based on making an hypothesis and getting feedback from the data that can validate or not that hypothesis. Also when using a distribution we get also measures on how confident we are about the estimated model. For example when estimating a Gaussian distribution, we estimate the coefficients μ and σ^2: the smallest the latter value the more confident we are about the model. When using a single value we loose this piece of information.

With the second method we get a density easy to work with but if the chosen density is very different from the posterior our model will not be very reliable.

The third method samples from the posterior and hence to get a good approximation but it can be difficult.

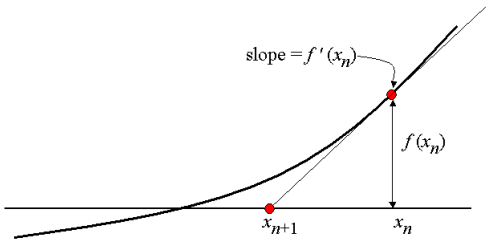


Figure 4: Newton-Raphsone example.

5.7.3 First method: point estimation, the MAP solution

We have seen that in some cases we cannot compute the posterior but we can compute the numerator of the expression, i.e., the prior multiplied by the likelihood. The value that maximizes the posterior is also the value that maximizes the numerator. We have already seen a likelihood maximization procedure in **??**; here we are maximizing the likelihood times the prior. This solution is the **maximum a posteriori (MAP)** estimate.

The Newton-Raphson method The Newton-Raphson method finds the points where $f(x) = 0$. Starting from an estimation x_n of such points, the estimation is updated by moving to the point where the tangent to the function at x_n passes throw the x-axis. This point is computed by approximating the gradient as a change in $f(x)$ divided by a change in x:

$$\begin{aligned} f'(x_n) &= -\frac{f(x_n) - 0}{x_n - x_{n+1}} \\ \Rightarrow x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \end{aligned} \tag{126}$$

Instead of finding the point for which $f(x) = 0$, we want to find the point where its derivative is 0. Hence we substitute $f(x)$ with $f'(x)$ and $f'(x)$ with $f''(x)$. In this way the method can be used to find points where the gradient passes through 0, i.e., minima, maxima and points of inflections:

$$\begin{aligned} f''(x_n) &= -\frac{f'(x_n) - 0}{x_n - x_{n+1}} \\ \Rightarrow x_{n+1} &= x_n - \frac{f'(x_n)}{f''(x_n)} \end{aligned} \tag{127}$$

When dealing with vectors, $f'(x)$ is replaced by the vector of partial derivatives evaluated at x_n and $\frac{1}{f''(x_n)}$ is replaced by the inverse of the Hessian matrix $\frac{-\partial^2 f(x)}{\partial \mathbf{x} \partial \mathbf{x}^T}$.

Derivation As already done, instead of maximizing the function itself, we maximize its logarithmic. Apart from the mathematical convenience this is also the advantage of avoiding underflow: we are dealing with probabilities, number between 0 and 1 and these might be too small to have a sufficient numerical precision. The logarithmic, as the numbers go to 0, makes the number goes to infinity, guaranteeing a better numerical precision. Since we cannot compute the derivative and set it to 0, we use the Newton-Raphson procedure. Let us call the numerator $g(\beta, \mathbf{X}, \mathbf{y}) = \Pr(\mathbf{y}|\beta, \mathbf{X})\Pr(\beta|\mathbf{X})$:

$$\beta' = \beta - \left(\frac{\partial^2 \log g(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \log g(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta} \tag{128}$$

The procedure is iterative and stops when the gradient is 0. To check the point we have converged to corresponds to a minimum we check the Hessian matrix is negative definite.

Before computing the derivatives we re-express $g(\beta, \mathbf{X}, \mathbf{y})$.

$$\begin{aligned} &\log g(\beta, \mathbf{X}, \mathbf{y}) = \sum_{n=1}^{N} \log \Pr(Y_n = y_n|\mathbf{x}_n, \beta) + \log \Pr(\beta|\sigma^2) = \\ &= \sum_{n=1}^{N} \log \left[\left(\frac{1}{1 + e^{-\beta^T \mathbf{x}_n}} \right)^{y_n} \left(\frac{e^{-\beta^T \mathbf{x}_n}}{1 + e^{-\beta^T \mathbf{x}_n}} \right)^{1-y_n} \right] + \log \Pr(\beta|\sigma^2) = \end{aligned} \tag{129}$$

We denote $P_n = P(Y_n = 1|\beta, \mathbf{x}_n) = \frac{1}{1+e^{-\beta^T \mathbf{x}_n}}$:

$$\log g(\beta, \mathbf{X}, \mathbf{y}) = \sum_{n=1}^{N} \log P_n^{y_n} + \log(1 - P_n)^{1-y_n} + \log \Pr(\beta|\sigma^2) = \tag{130}$$

Racalling

$$\begin{aligned}\Pr(\beta|\sigma^2) =& \frac{1}{\sqrt{(2\pi)^{\mathrm{D}}|\Sigma|}} \mathrm{e}^{-\frac{(\beta-\not{\mu})^{\mathrm{T}}(\beta-\not{\mu})}{2\sigma 2}} \\ \Rightarrow \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y}) =& \sum_{\mathrm{n}=1}^{\mathrm{N}} \log \mathrm{P}_{\mathrm{n}}^{\mathrm{y_n}} + \log(1-\mathrm{P_n})^{1-\mathrm{y_n}} - \frac{\mathrm{D}}{2}\log(2\pi) - \mathrm{D}\log\sigma + \\ & - \frac{1}{2\sigma^2}\beta^{\mathrm{T}}\beta\end{aligned} \tag{131}$$

Now we can take the derivative:

$$\begin{aligned}\frac{\partial \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta} =& -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\mathrm{y_n}}{\mathrm{P_n}} \frac{\partial \mathrm{P_n}}{\partial \beta} + \frac{1-\mathrm{y_n}}{1-\mathrm{P_n}} \frac{\partial(1-\mathrm{P_n})}{\partial \beta} = \\ =& \frac{\partial \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta} = -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\mathrm{y_n}}{\mathrm{P_n}} \frac{\partial \mathrm{P_n}}{\partial \beta} - \frac{1-\mathrm{y_n}}{1-\mathrm{P_n}} \frac{\partial \mathrm{P_n}}{\partial \beta}\end{aligned} \tag{132}$$

Now we must calculate $\frac{\partial \mathrm{P_n}}{\partial \beta}$:

$$\begin{aligned}\frac{\partial \mathrm{P_n}}{\partial \beta} =& \frac{\partial}{\partial \beta} \frac{1}{1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}} = \frac{\partial}{\partial \beta}\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right)^{-1} = \\ =& \frac{1}{\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right)^2} \frac{\partial}{\partial \beta}\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right) = \\ =& -\mathbf{x_n} \frac{\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}}{\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right)^2} = -\mathrm{x_n} \frac{1}{\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right)} \frac{\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}}{\left(1+\mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{x_n}}\right)} = \\ =& -\mathbf{x_n}\mathrm{P_n}(1-\mathrm{P_n})\end{aligned} \tag{133}$$

Substituting in 132:

$$\begin{aligned}\frac{\partial \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta} &= -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\mathrm{y_n}}{\mathrm{P_n}} \frac{\partial \mathrm{P_n}}{\partial \beta} - \frac{1-\mathrm{y_n}}{1-\mathrm{P_n}} \frac{\partial \mathrm{P_n}}{\partial \beta} = \\ &= -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\mathrm{y_n}}{\mathrm{P_n}} \left[-\mathbf{x_n}\mathrm{P_n}(1-\mathrm{P_n})\right] - \frac{1-\mathrm{y_n}}{1-\mathrm{P_n}} \left[-\mathbf{x_n}\mathrm{P_n}(1-\mathrm{P_n})\right] = \\ &= -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} -\mathrm{y_n}\mathbf{x_n}(1-\mathrm{P_n}) - (1-\mathrm{y_n})(-\mathbf{x_n}\mathrm{P_n}) = \\ &= -\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x_n}(\mathrm{y_n} - \mathrm{P_n})\end{aligned} \tag{134}$$

Now we must compute the Hessian matrix:

$$\begin{aligned}\frac{\partial^2 \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta \partial \beta^{\mathrm{T}}} &= \frac{\partial}{\partial \beta^{\mathrm{T}}} \left(-\frac{1}{\sigma^2}\beta + \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x_n}(\mathrm{y_n} - \mathrm{P_n}) \right) = -\frac{1}{\sigma^2}\mathbf{I} - \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x_n} \frac{\partial \mathrm{P_n}}{\partial \beta^{\mathrm{T}}} = \\ &= -\frac{1}{\sigma^2}\mathbf{I} - \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathbf{x_n}\mathbf{x_n^{\mathrm{T}}}\mathrm{P_n}(1-\mathrm{P_n})\end{aligned} \tag{135}$$

where we used the result from 133.

Note that the 2 terms are negative definite, hence the Hessian is always negative definite. Therefore there can only be one optimum and it must be the minimum.

The decision boundary is the one for which $\Pr(\mathrm{Y_n} = 1 | \mathbf{x}, \hat{\beta} = 0.5)$.

The steps above can be done for any prior and likelihood combination. In some cases the posterior might have several maxima and or some minima and it become difficult to know if the maximum is a global optimum.

The *Elements of Statistical Learning* book instead of maximizing posterior (or equivalently the numerator), it maximizes the likelihood. Let us express $\Pr(\mathrm{G} = \mathrm{k} | \mathrm{X} = \mathrm{x}) = \mathrm{p_k}(\mathrm{x}, \theta)$ with $\theta = \left\{\beta_{10}, \beta_1^{\mathrm{T}}, \cdots, \beta_{(\mathrm{K}-1)0}, \beta_{\mathrm{K}-1}^{\mathrm{T}}\right\}$ and let us define the log-likelihood

$$\mathrm{l}(\theta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \log \mathrm{p}_{\mathrm{g_i}}(\mathrm{x_i}, \theta) \tag{136}$$

Consider just two classes with responses 0, 1 and let $\mathrm{p}_1(\mathrm{x}, \theta) = \mathrm{p}(\mathrm{x}, \theta)$ and $\mathrm{p}_2(\mathrm{x}, \theta) = 1 - \mathrm{p}(\mathrm{x}, \theta)$. Recall that having two classes $\sum_{\mathrm{l}=1}^{\mathrm{K}-1} \mathrm{e}^{\beta_{\mathrm{l}0}+\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}_\mathrm{i}} = \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}$. The log-likelihood can be written as:

$$
\begin{aligned}
&\mathrm{l}(\beta) = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ \mathrm{y}_\mathrm{i} \log \mathrm{p}\left(\mathrm{x}_\mathrm{i}, \beta\right) + (1 - \mathrm{y}_\mathrm{i}) \log\left(1 - \mathrm{p}\left(\mathrm{x}_\mathrm{i}, \beta\right)\right) \right\} = \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ \mathrm{y}_\mathrm{i} \log \frac{\mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}{1 + \sum_{\mathrm{l}=1}^{\mathrm{K}-1} \mathrm{e}^{\beta_{\mathrm{l}0}+\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}_\mathrm{i}}} + (1 - \mathrm{y}_\mathrm{i}) \log\left(1 - \frac{\mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}{1 + \sum_{\mathrm{l}=1}^{\mathrm{K}-1} \mathrm{e}^{\beta_{\mathrm{l}0}+\beta_{\mathrm{l}}^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}\right) \right\} = \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ \mathrm{y}_\mathrm{i} \log \frac{\mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}{1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}} + (1 - \mathrm{y}_\mathrm{i}) \log\left(1 - \frac{\mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}{1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}}\right) \right\} = \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i} \left(\log \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}} - \log\left(1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}\right) \right) + (1 - \mathrm{y}_\mathrm{i}) \log \frac{1}{1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}} = \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i} \left(\log \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}} - \log\left(1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}\right) \right) - (1 - \mathrm{y}_\mathrm{i}) \log\left(1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}\right) = \\
&= \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{y}_\mathrm{i} \left(\beta_{10} + \beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}\right) - \log\left(1 + \mathrm{e}^{\beta_{10}+\beta_1^{\mathrm{T}}\mathrm{x}_\mathrm{i}}\right)
\end{aligned}
\tag{137}
$$

Here $\beta = [\beta_{10}, \beta_1]$. To maximize the log-likelihood we set the derivative to 0:

$$
\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}(\mathrm{y}_\mathrm{i} - \mathrm{p}(\mathrm{x}_\mathrm{i}, \beta) = 0 \tag{138}
$$

Using the *Newton-Raphson* algorithm that requires the second-derivative:

$$
\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} = -\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x}_\mathrm{i}\mathrm{x}_\mathrm{i}^{\mathrm{T}} \mathrm{p}(\mathrm{x}_\mathrm{i}, \beta)(1 - \mathrm{p}(\mathrm{x}_\mathrm{i}, \beta) \tag{139}
$$

$$
\Rightarrow \beta^{\mathrm{new}} = \beta^{\mathrm{old}} - \left(\frac{\partial^2 \ell\left(\beta\right)}{\partial \beta \partial \beta^{\mathrm{T}}} \right)^{-1} \frac{\partial \mathrm{l}(\beta)}{\partial \beta} \tag{140}
$$

Using the matrix notation,

$$\frac{\partial \ell(\beta)}{\partial \beta} = \mathrm{X}^{\mathrm{T}}(\mathrm{y} - \mathrm{p}) \tag{141}$$

$$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} = -\mathrm{X}^{\mathrm{T}}\mathrm{W}\mathrm{X} \tag{142}$$

So the Newton step becomes

$$\begin{aligned} \beta^{\text{new}} &= \beta^{\text{old}} + (\mathrm{X}^{\mathrm{T}}\mathrm{W}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}(\mathrm{y} - \mathrm{p}) = \\ &= (\mathrm{X}^{\mathrm{T}}\mathrm{W}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{W}\left(\mathrm{X}\beta^{\text{old}} + \mathrm{W}^{-1}(\mathrm{y} - \mathrm{p})\right) = \\ &= \left(\mathrm{X}^{\mathrm{T}}\mathrm{W}\mathrm{X}\right)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{W}\mathrm{z} \end{aligned} \tag{143}$$

with

$$\mathrm{z} = \mathrm{X}\beta^{\text{old}} + \mathrm{W}^{-1}(\mathrm{y} - \mathrm{p}) \tag{144}$$

sometimes known as the adjusted response. This algorithm is known as **iteratively reweighted least squares (IRLS)** since each iteration solves the weighted least square problem:

$$\beta^{\text{new}} \leftarrow \arg\min_{\beta} (\mathrm{z} - \mathrm{X}\beta)^{\mathrm{T}}\,\mathrm{W}\,(\mathrm{z} - \mathrm{X}\beta) \tag{145}$$

$\beta = 0$ seems a good starting value. Convergence is never guaranteed but typically the algorithm does converge, since the log-likelihood is concave but overshooting can occur. In the rare cases that the log-likelihood decreases, step size halving will guarantee convergence.

For $\mathrm{K} > 2$ we still use the iteration procedure but we will have a $\mathrm{K} - 1$ vector response and a non-diagonal weight matrix per observation. In this case it is better to work with the vector θ directly.

5.7.4 Second method: Laplace approximation

There are many approximation methods but the most common is the Laplace approximation. The idea is to approximate the density of interest with a Gaussian.

Choosing a Gaussian means choosing a proper variance and mean value. The Laplace approximation method fixes one of the two parameters, i.e., the

mean to the value maximizing the posterior, β. We approximate $\log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})$ with the Taylor expansion around $\hat{\beta}$:

$$\begin{aligned} \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y}) \approx \log \mathrm{g}(\hat{\beta}, \mathbf{X}, \mathbf{y}) + \cancel{\frac{\partial \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta}\Big|_{\hat{\beta}} \left(\beta - \hat{\beta}\right)} + \\ + \frac{\partial^2 \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta^2}\Big|_{\hat{\beta}} \left(\beta - \hat{\beta}\right)^2 = \\ = \log \mathrm{g}(\hat{\beta}, \mathbf{X}, \mathbf{y}) + \frac{\partial^2 \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta^2}\Big|_{\hat{\beta}} \frac{\left(\beta - \hat{\beta}\right)^2}{2} \end{aligned} \tag{146}$$

where the second term is the gradient evaluated at the maximum point that therefore must be 0.

The Gaussian density and its log are the following:

$$\begin{aligned} & \frac{1}{2\pi} \mathrm{e}^{-\frac{(\beta - \hat{\beta})^2}{2\sigma^2}} \\ & \Rightarrow \log \mathrm{const} - \frac{1}{2\sigma^2} (\beta - \mu)^2 \end{aligned} \tag{147}$$

which is similar to 146. By analogy of the two equations:

$$\begin{aligned} \mu &= \hat{\beta} \\ -\frac{1}{\sigma^2} &= -\frac{\partial^2 \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta^2}\Big|_{\hat{\beta}} \end{aligned} \tag{148}$$

It can be applied also to multivariate densities $\Pr(\beta, \mathbf{X}, \mathbf{y}) \approx \mathcal{N}(\mu, \mathbf{\Sigma})$:

$$\begin{aligned} \mu &= \hat{\beta} \\ \Sigma^{-1} &= -\frac{\partial^2 \log \mathrm{g}(\beta, \mathbf{X}, \mathbf{y})}{\partial \beta \partial \beta^{\mathrm{T}}}\Big|_{\hat{\beta}} \end{aligned} \tag{149}$$

Consider a multinomial 2D Gaussian function (dimensionaity of β is 2) as in 5 and suppose to project on the plane the points of the curve having the same value (the ellipses in the figure). The ellipses will be the combination of coefficients giving the same function value.

In figure 6 the Laplace approximation of the posterior (darker lines) is shown while the lighter lines are the true unnormalised posterior. The centre

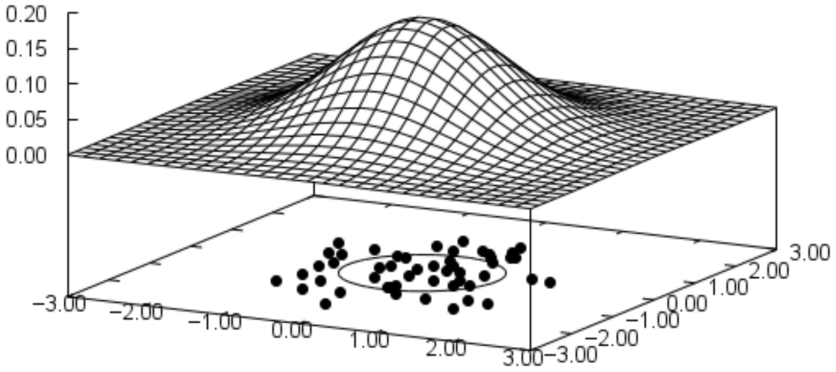


Figure 5: Example of multinomial Gaussian.

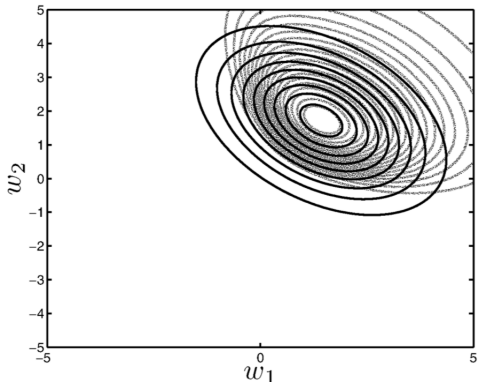


Figure 6: The axis are the parameters. The Laplace approximation of the posterior (darker lines) is shown while the lighter lines are the true unnormalised posterior. The centre point corresponds to the maximum point $\hat{\beta}$.

point corresponds to the maximum point $\hat{\beta}$. Note how the approximation is good around $\hat{\beta}$ and it diverges going further from it.

We use the approximate posterior to compute predictions. But now we have a density and not a single value: the prediction is computed by averaging over this density: it is like averaging over all possible values of β. We should calculate the expected value $\Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathbf{x}_{\mathrm{new}}, \beta)$ with respect to the approximate posterior denoted as $\mathcal{N}(\mu, \mathbf{\Sigma})$:

$$\Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathbf{x}_{\mathrm{new}}, \mathbf{X}, \beta) = \mathbf{E}_{\mathcal{N}(\mu,\mathbf{\Sigma})} \left\{\Pr(\mathrm{T}_{\mathrm{new}} = 1|\mathbf{x}_{\mathrm{new}}, \beta)\right\} \tag{150}$$

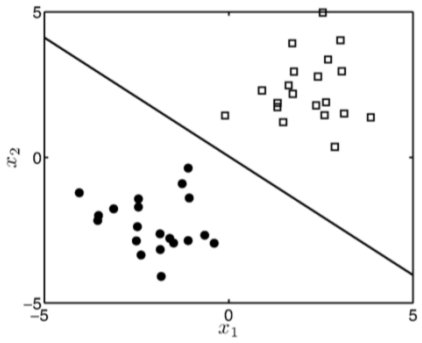
However, we cannot compute the integral (of the expectation), but we can sample from $\mathcal{N}(\mu, \mathbf{\Sigma})$ and approximate the expectation with a sum:

$$\Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathbf{x}_{\mathrm{new}}, \mathbf{X}, \beta) = \frac{1}{\mathrm{N_s}} \sum_{\mathrm{s}=1}^{\mathrm{N_s}} \frac{1}{1 + \mathrm{e}^{-\beta_{\mathrm{s}}^{\mathrm{T}} \mathbf{x}_{\mathrm{new}}}} \tag{151}$$

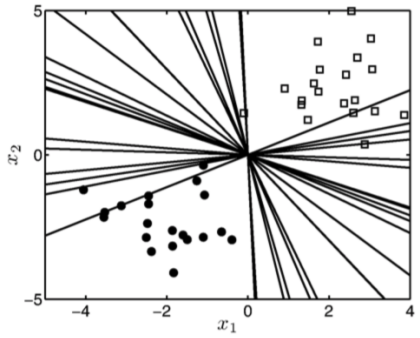
Comparison between the decision boundaries of MAP and Laplace method. In case of MAP estimation we get a single separating line as in 7a for the example of two classes. In case of Laplace approximation we don't have a single separating line but a density distribution. 7b shows 20 boundaries (i.e., set of coefficients) sampled randomly from the distribution. All or almost all separates the classes quite well but in the area from the graph further from the classes (or from the centroids or clusters) there is a lot of variability. This variability represents the uncertainty of the classifier in those area far from the classes, where no event (in the sense of data entry) was observed.

This is made clearer by looking at the decision boundaries in 7c and 7d. In case of MAP it is quite obvious: the probability increases or decreases just moving close to or far from to the boundary. In case of Laplace approximation the contours are no longer straight lines. The probability are now close to 0.5 in all the area except those closes to the two classes. It is like the classifier is unable to take a decision in those areas, that are the ones where no event has been noted.

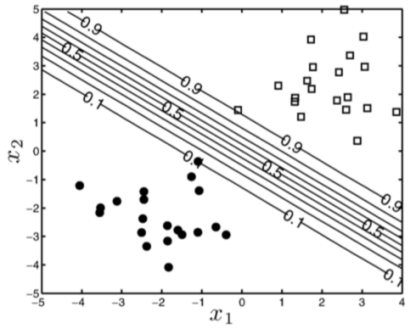
On the contrary the classifier resulting from MAP approximation is always sure about its work: it can only classify either with one or the other class except in on the points on the boundaries. This result from the fact that it is the result of a single point and not a distribution: i.e., the amount of confidence on the result is left out.



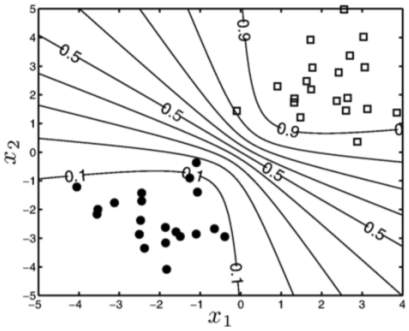
(a) Decision boundary for MAP estimation.



(b) Decision boundary for Laplace approximation estimation.



(c) Contours of probability of belonging to class 1 in case of MAP estimation.



(d) Contours of probability of belonging to class 1 in case of Laplace approximation estimation computed with a sample based approximation.

5.7.5 Third method: sampling technique

The reason of estimating the posterior density with Laplace approximation or any other method is to take into account uncertainties in β when making predictions. Using the Gaussian approximation we were able to sample as in equation 151 directly from data. When deciding we take the conditional probability by averaging all over the potential values β by taking the expectation:

$$\Pr(\mathrm{T}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \mathbf{X}, \mathbf{y}, \sigma^2) = \mathbf{E}_{\Pr(\beta|\mathbf{X},\mathbf{y},\sigma^2)}(\Pr(\mathrm{T}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \beta) \quad (152)$$

In these types of approximations we cut off the approximation step and sample directly from the posterior. A set of samples from the true posterior could be substituted directly into equation 151 to compute the desired predictive probability $\Pr(\mathrm{y_i}|\mathbf{x_i}, \mathbf{X}, \beta)$. A popular sampling technique is the **Metropolis-Hastings** algorithm.

Drawing from the posterior does not mean we can write it down but that we sample directly the system (physical or whatever) generating the data.

The metropolis-Hastings algorithms The objective is to sample from $\Pr(\beta|\mathbf{X}, \mathbf{y}, \sigma^2)$ to approximate the following expectation:

$$\begin{aligned} \Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \mathbf{X}, \mathbf{y}, \sigma^2) &= \mathbf{E}_{\Pr(\beta|\mathbf{X},\mathbf{y},\sigma^2)}\left[\Pr(\mathrm{T}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \beta)\right] \\ &= \int \Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \beta)\Pr(\beta|\mathbf{X}, \mathbf{y}, \sigma^2)\mathrm{d}\beta \end{aligned} \quad (153)$$

with

$$\Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \mathbf{X}, \mathbf{y}, \sigma^2) \approx \frac{1}{\mathrm{N_s}} \sum_{\mathrm{s}=1}^{\mathrm{N_s}} \Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathrm{x}_{\mathrm{new}}, \beta_{\mathrm{s}}) \quad (154)$$

The algorithm generates a sequence of samples $\beta_1, \beta_2, \cdots$. First all the algorithm is independent from the starting point, as long as we sample long enough: it is guaranteed the sequence converges.

The generation of new samples happens in this way. Suppose we have the sample $\mathrm{s}-1$, we will propose a new sample $\tilde{\beta}_{\mathrm{s}}$ and define the density

$\Pr(\tilde{\beta}_{\mathrm{s}}|\beta_{\mathrm{s}-1})$. This density is unrelated with the posterior $\Pr(\beta|\mathbf{X},\mathbf{y},\sigma^2)$ and we can define it as we please but it will affect the convergence time. A common choice is to use a Gaussian centred on the current sample, $\beta_{\mathrm{s}-1}$:

$$\Pr(\tilde{\beta}_{\mathrm{s}}|\beta_{\mathrm{s}-1},\Sigma) = \mathcal{N}(\beta_{\mathrm{s}-1},\Sigma) \tag{155}$$

Generally σ is taken diagonal with same values. The smaller the elements on the diagonal, the smaller the distance at each step.

Such a sequence creates a **random walk**. The choice of the Gaussian is justified bye the easy of sampling from the Gaussian and by its symmetry: moving from $\tilde{\beta}_{\mathrm{s}-1}$ to $\tilde{\beta}_{\mathrm{s}}$ is just as likely to move from $\tilde{\beta}_{\mathrm{s}}$ to $\tilde{\beta}_{\mathrm{s}-1}$:

$$\Pr(\tilde{\beta}_{\mathrm{s}}|\beta_{\mathrm{s}-1},\Sigma) = \Pr(\tilde{\beta}_{\mathrm{s}-1}|\beta_{\mathrm{s}},\Sigma) \tag{156}$$

Accepting or rejecting the candidate $\tilde{\beta}_{\mathrm{s}}$ is performed by calculating the following quantity:

$$\mathrm{r} = \frac{\Pr(\tilde{\beta}_{\mathrm{s}}|\mathbf{X},\mathbf{y},\sigma^2)}{\Pr(\beta_{\mathrm{s}-1}|\mathbf{X},\mathbf{y},\sigma^2)}\frac{\Pr(\beta_{\mathrm{s}-1}|\tilde{\beta}_{\mathrm{s}},\Sigma)}{\Pr(\hat{\beta}_{\mathrm{s}}|\beta_{\mathrm{s}-1},\Sigma)} \tag{157}$$

The expression is the product of the ratio of the posterior density at the proposed sample to that at the old sample times the ratio of the proposed densities. For the Gaussian symmetry discussed above, this term is 1 when using Gaussian densities. Note that we cannot compute exactly the densities, but being a ratio the normalisation constant (the denominator in Bayes expression) simplifies and only the likelihoods times the priors are left:

$$\mathrm{r} = \frac{\Pr(\tilde{\beta}_{\mathrm{s}}|\sigma^2)}{\Pr(\beta_{\mathrm{s}-1}|\sigma^2)}\frac{\Pr(\mathbf{y}|\mathbf{X},\tilde{\beta}_{\mathrm{s}},\sigma^2)}{\Pr(\mathbf{y}|\mathbf{X},\beta_{\mathrm{s}-1},\sigma^2)} \tag{158}$$

If $\mathrm{r} > 1$, i.e, we get a higher posterior density, we accept the candidate otherwise we accept the candidate with probability equal to r. The complete algorithm is depicted in 8 where a uniform distribution in $[0,1]$ is used as decision rule in case $\mathrm{r} < 1$. Being a uniform distribution, the probability that $\mathrm{u} \leq \mathrm{r}$ is r.

9 shows a 10-iteration process of the algorithms where solid lines are accepted coefficients, dashed lines are the rejected ones. Note that even the third sample causes a decrease in the posterior $\mathrm{r} < 1$, nevertheless in this specific case the decision rule accepted it. On the contrary the 4-th

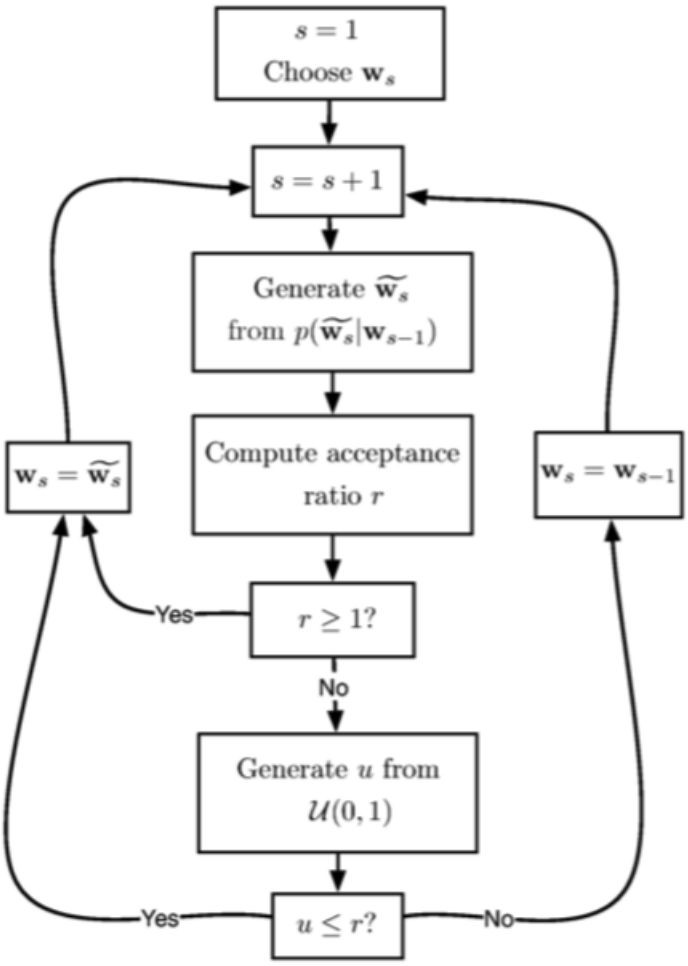
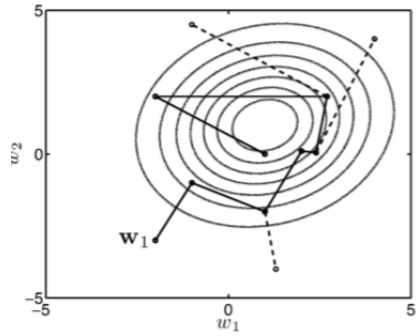


Figure 8: The steps of the Metropolis-Hastings algorithm. U is a uniform distribution in $[0, 1]$



(e) After ten samples.

Figure 9: Example of Metropolis-Hastings iterations: solid lines are accepted coefficients, dashed lines are the rejected ones.

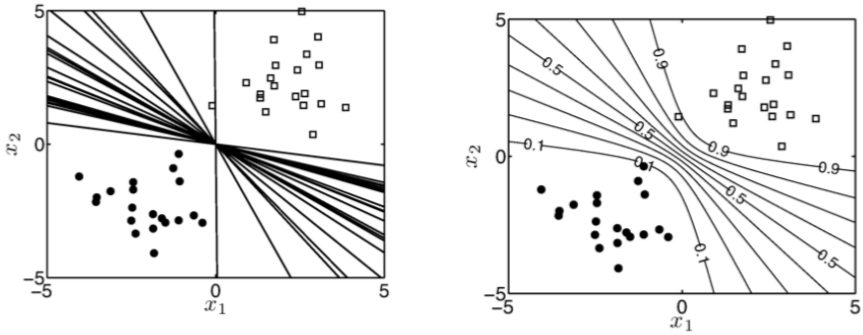
sample causes a huge decrease hence it is very unlikely it is accepted and in fact it is not, so $\beta_4 = \beta_3$. After N samples we compute the sample based approximations to the mean and covariance

$$\begin{aligned} \mu' &= \frac{1}{\mathrm{N_s}} \sum_{\mathrm{s}=1}^{\mathrm{N_s}} \beta_{\mathrm{s}} \\ \mathrm{S}' &= \frac{1}{\mathrm{N_s}} \sum_{\mathrm{s}=1}^{\mathrm{N_s}} (\beta_{\mathrm{s}} - \mu')(\beta_{\mathrm{s}} - \mu')^{\mathrm{T}} \end{aligned} \tag{159}$$

Definition 5.1. Burn-in It is the time interval between the starting of the algorithm and the convergence.

It cannot be determined: to overcome this problem, a method for determining convergence (to a distribution, not to a value) should be established. For example, in the algorithm above, we do not know if the starting point belongs to an area from which we are supposed to sample: it might be very far from the posterior. Including these samples in the approximation might result in a not good value. That is why the first samples (ranging from few samples to thousands) should be discarded.

A popular method is to start several samplers simultaneously from different starting points. When all the samplers are generating samples with similar mean and variance, it suggests they converged all to the same distributions.



(a) Predictive probability contours showing the probability of classifying objects at any location as square.

(b) Decision boundaries created from randomly selected MH samples.

Figure 10: Example of Metropolis-Hastings algorithm results.

Definition 5.2. convergence In this case we are talking about the convergence to a given distribution, not a single point. The convergence to a distribution is characterized to the convergence of its parameters: in case of the Gaussian the mean and variance.

We can even look at each coefficient independently:

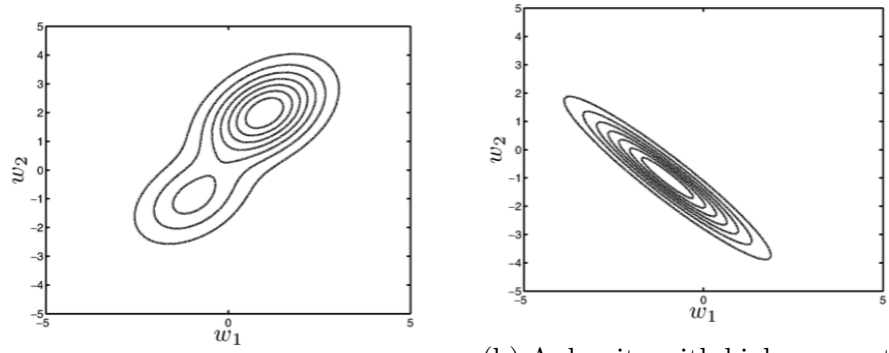
$$\Pr(\beta_1|\mathbf{X},\mathbf{y},\sigma^2) = \int \Pr(\beta_1,\beta_2|\mathbf{X},\mathbf{y},\sigma^2)\mathrm{d}\beta_2 \tag{160}$$

To calculate the predictive probability using the obtained set of samples, we can do what already done with the Laplace approximation:

$$\Pr(\mathrm{Y}_{\mathrm{new}} = 1|\mathbf{x}_{\mathrm{new}},\mathbf{X},\mathbf{y},\sigma^2) = \frac{1}{\mathrm{N_s}}\sum_{\mathrm{s}=1}^{\mathrm{N_s}} \frac{1}{1+\mathrm{e}^{-\beta_{\mathrm{s}}^{\mathrm{T}}\mathbf{x}_{\mathrm{new}}}} \tag{161}$$

10a and 10b show an example of possible shapes of contours: it does not look too different from the Laplace ones. The only difference is that these contours are not so tight as the Laplace's ones. This suggests the probability decreases more slowly.

Limitations The difficult mostly lies in the unknown shape of the density. When a density has two or more modes, MH moves towards the modes as



(a) A bi-modal density.

(b) A density with high parameter correlation.

Figure 11: Example of Metropolis-Hastings algorithm results.

these moves increase the posterior density and hence are always accepted. When close to a mode, many steps are required to move from one mode to another and this is very unlikely. We might end up exploring a mode without even knowing the other exists (see 11a). Another problem arises when the variables are strongly correlated (see 11b). Let us pick any position and propose a movement from a Gaussian with diagonal covariance (i.e., having circular contours). The shapes are very different and many samples will be rejected: the majority of moves that we sample from our proposal will involve moving steeply down the probability gradient. There are even other problems.

5.7.6 Usage

LR is used as data analysis tool where the goal is to understand the role of the input variables in explaining the outcome. Typically many models are fit in a search for a parsimonious model involving a subset of the variables, possibly with some interactions terms.

It is widely used in biostatistical applications where binary responses (two classes) occur quite frequently. For example, patients survive or die, have heart disease or not, or a condition is present or absent.

5.8 Regularized Logistic regression

We can use the L_1 penalty for variable selection and shrinkage:

$$\underset{\beta_0,\beta_1}{\arg\max} \left\{ \sum_{i=1}^{N} y_i \left(\beta_0 + \beta^T x_i\right) - \log\left(1 + e^{\beta_0+\beta^T x_i}\right) - \lambda \sum_{j=-1}^{p} |\beta_j| \right\} \quad (162)$$

This function is concave and can be solved using a nonlinear programming method.

5.9 Logistic vs LDA

The difference between the models relies on how the linear coefficients are estimated. The logistic regression model is more general since it makes less assumptions.

LDA is not robust to outliers since observations far from the decision boundary are used to estimate the common covariance matrix, while they are scaled down in the Logistic regression.

5.10 Perceptron learning algorithm

It tries to find a separate hyperplane by minimizing the distance of misclassified points to the decision boundary. If a response $y_i = 1$ is misclassified, then $x_i^T\beta+\beta_0 < 0$, and the opposite for a misclassified response with $y_i = -1$. The goal is to minimize

$$D(\beta, \beta_0) = -\sum_{i\in\mathcal{M}} y_i \left(x_i^T\beta + \beta_0\right) \quad (163)$$

where $\mathcal{M}$ is the set of misclassified points. The gradient is

$$\frac{\partial D(\beta, \beta_0}{\partial\beta} = -\sum_{i\in\mathcal{M}} y_i x_i \quad (164)$$

$$\frac{\partial D(\beta, \beta_0}{\partial\beta_0} = -\sum_{i\in\mathcal{M}} y_i \quad (165)$$

where the algorithm uses the stochastic gradient descent where the coefficients are updated by the gradient value weighted by a step ρ. There are many problems though:

- when data are separable, there are many solutions which depend on the starting value;
- many steps might be required;
- when data are not separable, the algorithm will not converge;

5.11 Optimal separating hyperplanes

The optimal separating hyperplane separates the two classes and maximizes the distance to the closest point from either class (Vapnik, 1996). Not only does this provide a unique solution to the separating hyperplane problem, but by maximizing the margin between the two classes on the training data, this leads to better classification performance on test data.

Suppose M is the distance. Then

$$\begin{aligned} &\max_{\beta,\beta_0,||\beta||=1} \mathrm{M} \\ &\mathrm{y_i}(\mathrm{x_i^T}\beta + \beta_0) \geq \mathrm{M} \end{aligned} \tag{166}$$

We can move the constrain on the module $|||\beta|| = 1$ to the condition $\frac{1}{||\beta||}\mathrm{y_i}(\mathrm{x_i^T}\beta + \beta_0) \geq \mathrm{M}$, which redefines β_0.

For a pair of coefficients satisfying this inequalities, any positively scaled multiple satisfies them too, so we can set arbitrarily set $||\beta|| = \frac{1}{\mathrm{M}}$:

$$\begin{aligned} &\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 \\ &\textit{subject to } \mathrm{y_i}(\mathrm{x_i^T}\beta + \beta_0) \geq 1 \end{aligned} \tag{167}$$

With this constraint, a margin around the decision boundary with thick $\frac{1}{||\beta||}$ is present. We choose β to maximize the thickness of margin margin. This is a convex optimization problem. The Lagrange primal function to be minimized is:

$$\mathrm{L_p} = \frac{1}{2}||\beta||^2 = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_\mathrm{i} \left[\mathrm{y_i}\left(\mathrm{x_i^T}\beta + \beta_0\right)\right] \tag{168}$$

and setting the derivatives to 0:

$$\beta = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_\mathrm{i}\mathrm{y_i}\mathrm{x_i}0 \quad = \sum_{\mathrm{i}=1}^{\mathrm{N}} \alpha_\mathrm{i}\mathrm{y_i} \tag{169}$$

and substituting in 168:

$$\begin{aligned} L_D &= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k \\ &\textit{subject to } \alpha_i \geq 0 \textit{ and } \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned} \tag{170}$$

The optimal hyperplane focuses more on points that counts, i.e., close to the border. For this reason it is more robust. LDA depends on all the data, even points faraway. If classes are really Gaussian LDA is optimal and separating hyperplane will pay a price for focusing on noisier data. For logistic regression, if a separate hyperplane exists, since the log-likelihood can be driven to 0.

When data are not separable there wil be no feasible solution and an alternative formulation is needed.