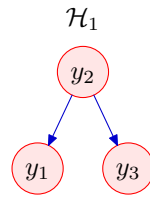


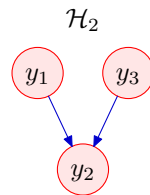
## Gaussian Quiz



1. Assuming that the variables  $y_1, y_2, y_3$  in this belief network have a joint Gaussian distribution, which of the following matrices *could be* the covariance matrix?

A	B	C	D
$\begin{bmatrix} 9 & 3 & 1 \\ 3 & 9 & 3 \\ 1 & 3 & 9 \end{bmatrix}$	$\begin{bmatrix} 8 & -3 & 1 \\ -3 & 9 & -3 \\ 1 & -3 & 8 \end{bmatrix}$	$\begin{bmatrix} 9 & 3 & 0 \\ 3 & 9 & 3 \\ 0 & 3 & 9 \end{bmatrix}$	$\begin{bmatrix} 9 & -3 & 0 \\ -3 & 10 & -3 \\ 0 & -3 & 9 \end{bmatrix}$

2. Which of the matrices could be the *inverse* covariance matrix?



3. Which of the matrices could be the covariance matrix of the second graphical model?

4. Which of the matrices could be the inverse covariance matrix of the second graphical model?

5. Let three variables  $y_1, y_2, y_3$  have covariance matrix  $\mathbf{K}_{(3)}$ , and inverse covariance matrix  $\mathbf{K}_{(3)}^{-1}$ .

$$\mathbf{K}_{(3)} = \begin{bmatrix} 1 & .5 & 0 \\ .5 & 1 & .5 \\ 0 & .5 & 1 \end{bmatrix} \quad \mathbf{K}_{(3)}^{-1} = \begin{bmatrix} 1.5 & -1 & .5 \\ -1 & 2 & -1 \\ .5 & -1 & 1.5 \end{bmatrix}$$

Now focus on the variables  $y_1$  and  $y_2$ . Which statements about *their* covariance matrix  $\mathbf{K}_{(2)}$  and inverse covariance matrix  $\mathbf{K}_{(2)}^{-1}$  are true?

(A)	(B)
$\mathbf{K}_{(2)} = \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}$	$\mathbf{K}_{(2)}^{-1} = \begin{bmatrix} 1.5 & -1 \\ -1 & 2 \end{bmatrix}$

# The Humble Gaussian Distribution

David J.C. MacKay  
Cavendish Laboratory  
Cambridge CB3 0HE  
United Kingdom

June 11, 2006 – Draft 1.0

## Abstract

These are elementary notes on Gaussian distributions, aimed at people who are about to learn about Gaussian processes. I emphasize the following points.

What happens to a covariance matrix and inverse covariance matrix when we omit a variable.

What it means to have zeros in a covariance matrix.

What it means to have zeros in an inverse covariance matrix.

How probabilistic models expressed in terms of ‘energies’ relate to Gaussians.

Why eigenvectors and eigenvalues don’t have any fundamental status.

## 1 Introduction

Let’s chat about a Gaussian distribution with zero mean, such as

$$P(\mathbf{y}) = \frac{1}{Z} e^{-\frac{1}{2} \mathbf{y}^\top \mathbf{A} \mathbf{y}}, \quad (1)$$

where  $\mathbf{A} = \mathbf{K}^{-1}$  is the inverse of the covariance matrix,  $\mathbf{K}$ . I’m going to emphasize dimensions throughout this note, because I think dimension-consciousness enhances understanding.<sup>1</sup> I’ll write

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{bmatrix} \quad (4)$$

---

<sup>1</sup>It’s conventional to write the diagonal elements in  $\mathbf{K}$  as  $\sigma_i^2$  and the offdiagonal elements as  $\sigma_{ij}$ . For example

$$\mathbf{K} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \quad (2)$$

A confusing convention, since it implies that  $\sigma_{ij}$  has different dimensions from  $\sigma_i$ , even if all axes  $i, j$  have the same dimensions!

Another way of writing an off-diagonal coefficient is

$$K_{ij} = \rho_{ij} \sigma_i \sigma_j, \quad (3)$$

where  $\rho$  is the correlation coefficient between  $i$  and  $j$ . This is a better notation since it’s dimensionally consistent in the way it uses the letter  $\sigma$ . But I will stick with the notation  $K_{ij}$ .

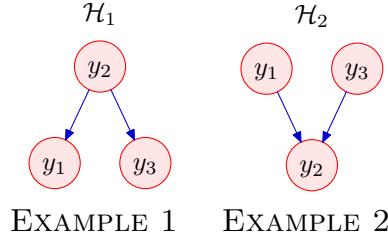
The definition of the covariance matrix is

$$K_{ij} = \langle y_i y_j \rangle \quad (5)$$

so the dimensions of the element  $K_{ij}$  are (dimensions of  $y_i$ ) times (dimensions of  $y_j$ ).

## 1.1 Examples

Let's work through a few graphical models.



### 1.1.1 Example 1

Maybe  $y_2$  is the temperature outside some buildings (or rather, the deviation of the outside temperature from its mean), and  $y_1$  is the temperature deviation inside building 1, and  $y_3$  is the temperature inside building 3. This graphical model says that if you know the outside temperature  $y_2$  then  $y_1$  and  $y_3$  are independent.

Let's consider this generative model:

$$y_2 = \nu_2 \quad (6)$$

$$y_1 = w_1 y_2 + \nu_1 \quad (7)$$

$$y_3 = w_3 y_2 + \nu_3, \quad (8)$$

where  $\{\nu_i\}$  are independent normal variables with variances  $\{\sigma_i^2\}$ .

Then we can write down the entries in the covariance matrix, starting with the diagonal entries

$$K_{11} = \langle y_1 y_1 \rangle = \langle (w_1 \nu_2 + \nu_1)(w_1 \nu_2 + \nu_1) \rangle = w_1^2 \langle \nu_2^2 \rangle + 2w_1 \langle \nu_1 \nu_2 \rangle + \langle \nu_1^2 \rangle = w_1^2 \sigma_2^2 + \sigma_1^2 \quad (9)$$

$$K_{22} = \sigma_2^2 \quad (10)$$

$$K_{33} = w_3^2 \sigma_2^2 + \sigma_3^2 \quad (11)$$

So we can fill in this much:

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{bmatrix} = \begin{bmatrix} w_1^2 \sigma_2^2 + \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & w_3^2 \sigma_2^2 + \sigma_3^2 \end{bmatrix} \quad (12)$$

The off diagonal terms are

$$K_{12} = \langle y_1 y_2 \rangle = \langle (w_1 \nu_2 + \nu_1)(\nu_2) \rangle = w_1 \sigma_2^2 \quad (13)$$

(and similarly for  $K_{23}$ ) and

$$K_{13} = \langle y_1 y_3 \rangle = \langle (w_1 \nu_2 + \nu_1)(w_3 \nu_2 + \nu_3) \rangle = w_1 w_3 \sigma_2^2 \quad (14)$$

So the covariance matrix is:

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{bmatrix} = \begin{bmatrix} w_1^2 \sigma_2^2 + \sigma_1^2 & w_1 \sigma_2^2 & w_1 w_3 \sigma_2^2 \\ & \sigma_2^2 & w_3 \sigma_2^2 \\ & & w_3^2 \sigma_2^2 + \sigma_3^2 \end{bmatrix} \quad (15)$$

(where the remaining blank elements can be filled in by symmetry).

Now let's think about the inverse covariance matrix.

One way to get to it is to write down the joint distribution.

$$P(y_1, y_2, y_3 | \mathcal{H}_1) = P(y_2)P(y_1 | y_2)P(y_3 | y_2) \quad (16)$$

$$= \frac{1}{Z_2} \exp\left(-\frac{y_2^2}{2\sigma_2^2}\right) \frac{1}{Z_1} \exp\left(-\frac{(y_1 - w_1 y_2)^2}{2\sigma_1^2}\right) \frac{1}{Z_3} \exp\left(-\frac{(y_3 - w_3 y_2)^2}{2\sigma_3^2}\right) \quad (17)$$

We can now collect all the terms in  $y_i y_j$ .

$$\begin{aligned} P(y_1, y_2, y_3) &= \frac{1}{Z'} \exp\left(-\frac{y_2^2}{2\sigma_2^2} - \frac{(y_1 - w_1 y_2)^2}{2\sigma_1^2} - \frac{(y_3 - w_3 y_2)^2}{2\sigma_3^2}\right) \\ &= \frac{1}{Z'} \exp\left(-y_2^2 \left[\frac{1}{2\sigma_2^2} + \frac{w_1^2}{2\sigma_1^2} + \frac{w_3^2}{2\sigma_3^2}\right] - y_1^2 \frac{1}{2\sigma_1^2} + 2y_1 y_2 \frac{w_1}{2\sigma_1^2} - y_3^2 \frac{1}{2\sigma_3^2} + 2y_3 y_2 \frac{w_3}{2\sigma_3^2}\right) \\ &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left[\frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2}\right] & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}\right) \end{aligned}$$

So the inverse covariance matrix is

$$\mathbf{K}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left[\frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2}\right] & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix}$$

信息矩阵里的0意味着如果其他元素确定情况下，i和j元素相对独立

The first thing I'd like you to notice here is the zeroes.  $[\mathbf{K}^{-1}]_{13} = 0$ .

The meaning of a zero in an inverse covariance matrix (at location  $i, j$ ) is *conditional on all the other variables, these two variables  $i$  and  $j$  are independent*.

Next, notice that whereas  $y_1$  and  $y_2$  were positively correlated (assuming  $w_1 > 0$ ), the coefficient  $[\mathbf{K}^{-1}]_{12}$  is negative. It's common that a covariance matrix  $\mathbf{K}$  in which all the elements are non-negative has an inverse that includes some negative elements. So positive off-diagonal terms in the covariance matrix always describe positive correlation; but the off-diagonal terms in the inverse covariance matrix can't be interpreted that way. The sign of an element  $(i, j)$  in the inverse covariance matrix does *not* tell you about the correlation between those two variables. For example, remember: there is a zero at  $[\mathbf{K}^{-1}]_{13}$ . But that doesn't mean that variables  $y_1$  and  $y_3$  are uncorrelated. Thanks to their parent  $y_2$ , they are correlated, with covariance  $w_1 w_3 \sigma_2^2$ .

The off-diagonal entry  $[\mathbf{K}^{-1}]_{ij}$  in an inverse covariance matrix indicates *how  $y_i$  and  $y_j$  are correlated if we condition on all the other variables apart from those two*: if  $[\mathbf{K}^{-1}]_{ij} < 0$ , they are *positively* correlated, conditioned on the others; if  $[\mathbf{K}^{-1}]_{ij} > 0$ , they are *negatively* correlated.

信息矩阵中如果某个位置的元素为正，则说明i和j元素在其他元素确定情况下呈负相关；反之呈正相关；如果=0，则相互独立。

信息矩阵可以读出条件概率的信息，读出所有其他变量确定情况下的条件概率

The inverse covariance matrix is great for reading out properties of conditional distributions in which we condition on *all the variables except one*.

For example, look at  $[\mathbf{K}^{-1}]_{11} = \frac{1}{\sigma_1^2}$ ; if we know  $y_2$  and  $y_3$ , then the probability distribution of  $y_1$  is Gaussian with variance  $1/[\mathbf{K}^{-1}]_{11}$ . That one was easy.

Look at  $[\mathbf{K}^{-1}]_{22} = \left[ \frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2} \right]$ . if we know  $y_1$  and  $y_3$ , then the probability distribution of  $y_2$  is Gaussian with variance

$$\frac{1}{[\mathbf{K}^{-1}]_{22}} = \frac{1}{\frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2}}. \quad (18)$$

协方差的逆很有意思，y2协方差仅有自身，而其逆加入了先验和似然，这显然使得分母变大了，提高了精度，这一点和贝叶斯推论一致

That's not so obvious, but it's familiar if you've applied Bayes theorem to Gaussians – when we do inference of a parent like  $y_2$  given its children, the inverse-variances of the prior and the likelihoods add. Here, the parent variable's inverse variance (also known as its precision) is the sum of the precision contributed by the prior  $\frac{1}{\sigma_2^2}$ , the precision contributed by the measurement of  $y_1$ ,  $\frac{w_1^2}{\sigma_1^2}$ , and the precision contributed by the measurement of  $y_3$ ,  $\frac{w_3^2}{\sigma_3^2}$ .

The off-diagonal entries in  $\mathbf{K}$  tell us how the *mean* of [the conditional distribution of one variable given the others] depends on [the others].

Let's take variable  $y_3$  conditioned on the other two, for example.

$$\begin{aligned} P(y_3 | y_1, y_2, \mathcal{H}_1) &\propto P(y_1, y_2, y_3 | \mathcal{H}_1) \\ &\propto \frac{1}{Z'} \exp \left( -\frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left[ \frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2} \right] & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \right) \end{aligned}$$

Let's highlight in Blue the terms  $y_1, y_2$  that are fixed and known and uninteresting, and highlight in Green everything that is multiplying the interesting term  $y_3$ .

$$\begin{aligned} P(y_3 | y_1, y_2, \mathcal{H}_1) &\propto P(y_1, y_2, y_3 | \mathcal{H}_1) \\ &\propto \frac{1}{Z'} \exp \left( -\frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left[ \frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2} \right] & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \right) \end{aligned}$$

All those blue multipliers in the central matrix aren't achieving anything. We can just ignore them (and redefine the constant of proportionality). For the benefit of anyone with a colour-blind printer, here it is again:

$$P(y_3 | y_1, y_2, \mathcal{H}_1) \propto \exp \left( -\frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \right)$$

$$P(y_3 | y_1, y_2, \mathcal{H}_1) \propto \exp \left( -\frac{1}{2} \frac{1}{\sigma_3^2} [y_3]^2 - [y_3] \left[ 0 \times y_1 - \frac{w_3}{\sigma_3^2} \times y_2 \right] \right)$$

We obtain the mean by completing the square.<sup>2</sup>

$$P(y_3 | y_1, y_2, \mathcal{H}_1) \propto \exp \left( -\frac{1}{2} \frac{1}{\sigma_3^2} \left[ y_3 - \frac{\left[ 0 \times y_1 + \frac{w_3}{\sigma_3^2} \times y_2 \right]}{\frac{1}{\sigma_3^2}} \right]^2 \right)$$

In this case, this all collapses down, of course, to

$$P(y_3 | y_1, y_2, \mathcal{H}_1) \propto \exp \left( -\frac{1}{2} \frac{1}{\sigma_3^2} [y_3 - w_3 y_2]^2 \right), \quad (19)$$

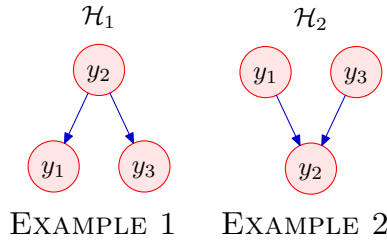
as defined in the original generative model (8).

In general, the offdiagonal coefficients  $K^{-1}$  tell us the sensitivity of [the mean of the conditional distribution] to the other variables.

$$\mu_{y_3 | y_1, y_2} = \frac{-K_{13}^{-1} y_1 - K_{23}^{-1} y_2}{K_{33}^{-1}} \quad (20)$$

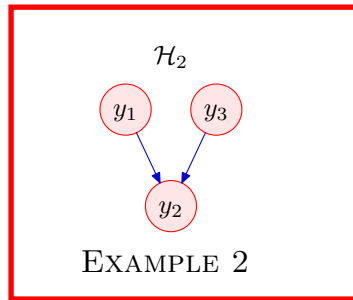
信息矩阵展示了一个量  
依靠其他确定量的均值  
是信息矩阵对应位置的  
线性组合

So the conditional mean of  $y_3$  is a linear function of the **known variables**, and the offdiagonal entries in  $K^{-1}$  tell us the *coefficients* in that linear function.



## 1.2 Example 2

Here's another example, where two parents have one child. For example, the price of electricity  $y_2$  from a power station might depend on the price of gas,  $y_1$ , and the price of carbon emission rights,  $y_3$ .



这个结构刚好是例1的对  
偶问题，那么它的协方  
差矩阵应该与例子1中的  
信息矩阵有相同的结构

<sup>2</sup>'Completing the square' is  $\frac{1}{2}ay^2 - by = \frac{1}{2}a(y - b/a)^2 + \text{constant}$ .

$$y_2 = w_1 y_1 + w_3 y_3 + \nu_2 \quad (21)$$

Note that the units in which gas price, electricity price, and carbon price are measured are all different (pounds per cubic metre, pennies per kWh, and euros per tonne, for example). So  $y_1$ ,  $y_3$ , and  $y_3$  have different dimensions from each other. Most people who do data modelling treat their data as ‘just numbers’, but I think it is a useful discipline to keep track of dimensions and to carry out only dimensionally valid operations. [Dimensionally valid operations satisfy the two rules of dimensions: (1) only add, subtract and compare quantities that have like dimensions; (2) arguments of all functions like exp, log, sin must be dimensionless. Rule 2 is really just a special case of rule 1, since  $\exp(x) = 1 + x + x^2 + \dots$ , so to satisfy rule 1, the dimensions of  $x$  must be the same as the dimensions of 1.]

What is the covariance matrix?

Here we assume that the parent variables  $y_1$  and  $y_3$  are uncorrelated.

The covariance matrix is

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & w_1 \sigma_1^2 & 0 \\ \sigma_2^2 + w_1^2 \sigma_1^2 + w_3^2 \sigma_3^2 & w_3 \sigma_3^2 \\ \sigma_3^2 \end{bmatrix} \quad (22)$$

Notice the zero correlation between the uncorrelated variables (1, 3).

What do you think the (1, 3) entry in the inverse covariance matrix will be? Let’s work it out in the same way as before. The joint distribution is

$$P(y_1, y_2, y_3 | \mathcal{H}_2) = P(y_1)P(y_3)P(y_2 | y_1, y_3) \quad (23)$$

$$= \frac{1}{Z_1} \exp\left(-\frac{y_1^2}{2\sigma_1^2}\right) \frac{1}{Z_3} \exp\left(-\frac{y_3^2}{2\sigma_3^2}\right) \frac{1}{Z_2} \exp\left(-\frac{(y_2 - w_1 y_1 - w_3 y_3)^2}{2\sigma_2^2}\right) \quad (24)$$

We collect all the terms in  $y_i y_j$ .

$$\begin{aligned} P(y_1, y_2, y_3) &= \frac{1}{Z'} \exp\left(-\frac{y_1^2}{2\sigma_1^2} - \frac{y_3^2}{2\sigma_3^2} - \frac{(y_2 - w_1 y_1 - w_3 y_3)^2}{2\sigma_2^2}\right) \\ &= \frac{1}{Z'} \exp\left(-y_1^2 \left[\frac{1}{2\sigma_1^2} + \frac{w_1^2}{2\sigma_2^2}\right] - y_2^2 \frac{1}{2\sigma_2^2} + 2y_1 y_2 \frac{w_1}{2\sigma_1^2} \right. \\ &\quad \left. - y_3^2 \left[\frac{1}{2\sigma_3^2} + \frac{w_3^2}{2\sigma_2^2}\right] + 2y_3 y_2 \frac{w_3}{2\sigma_2^2} - 2y_3 y_1 \frac{w_1 w_3}{2\sigma_2^2}\right) \\ &= \frac{1}{Z'} \exp\left(-\frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} \left[\frac{1}{2\sigma_1^2} + \frac{w_1^2}{\sigma_2^2}\right] & -\frac{w_1}{\sigma_2^2} & +\frac{w_1 w_3}{\sigma_2^2} \\ -\frac{w_1}{\sigma_2^2} & \frac{1}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} \\ +\frac{w_1 w_3}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} & \left[\frac{1}{2\sigma_3^2} + \frac{w_3^2}{\sigma_2^2}\right] \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}\right) \end{aligned}$$

So the inverse covariance matrix is

$$\mathbf{K}^{-1} = \begin{bmatrix} \left[\frac{1}{2\sigma_1^2} + \frac{w_1^2}{\sigma_2^2}\right] & -\frac{w_1}{\sigma_2^2} & +\frac{w_1 w_3}{\sigma_2^2} \\ -\frac{w_1}{\sigma_2^2} & \frac{1}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} \\ +\frac{w_1 w_3}{\sigma_2^2} & -\frac{w_3}{\sigma_2^2} & \left[\frac{1}{2\sigma_3^2} + \frac{w_3^2}{\sigma_2^2}\right] \end{bmatrix} \quad (25)$$

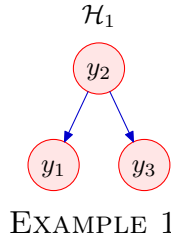
<1,2>为负表明他们是正相关的

Notice (assuming  $w_1 > 0$  and  $w_3 > 0$ ) that the offdiagonal term connecting a parent and a child  $[\mathbf{K}^{-1}]_{12}$  is negative and the offdiagonal term connecting the two parents  $[\mathbf{K}^{-1}]_{13}$  is positive. This positive term indicates that, conditional on all the other variables (i.e.,  $y_2$ ), the two parents  $y_1$  and  $y_3$  are anticorrelated. That's 'explaining away'. Once you know the price of electricity was average, for example, you can deduce that if gas was more expensive than normal, carbon probably was less expensive than normal.

## 2 Omission of one variable

<1,3>位置表示两个父类的相关性，为正表明负相关，可以着吗解释：如果y2确定，那么y1和y3肯定是此消彼长的关系，所以是负相关

Consider example 1.



The covariance matrix of all three variables is:

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{12} & K_{22} & K_{23} \\ K_{13} & K_{23} & K_{33} \end{bmatrix} = \begin{bmatrix} w_1^2 \sigma_2^2 + \sigma_1^2 & w_1 \sigma_2^2 & w_1 w_3 \sigma_2^2 \\ w_1 \sigma_2^2 & \sigma_2^2 & w_3 \sigma_2^2 \\ w_1 w_3 \sigma_2^2 & w_3 \sigma_2^2 & w_3^2 \sigma_2^2 + \sigma_3^2 \end{bmatrix} \quad (26)$$

If we decide we want to talk about the joint distribution of just  $y_1$  and  $y_2$ , the covariance matrix is simply the sub-matrix:

$$\mathbf{K}_2 = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix} = \begin{bmatrix} w_1^2 \sigma_2^2 + \sigma_1^2 & w_1 \sigma_2^2 \\ w_1 \sigma_2^2 & \sigma_2^2 \end{bmatrix} \quad (27)$$

This follows from the definition of the covariance,

$$K_{ij} = \langle y_i y_j \rangle. \quad (28)$$

The inverse covariance matrix, on the other hand, does not change in such a simple way. The  $3 \times 3$  inverse covariance matrix was:

$$\mathbf{K}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} & 0 \\ -\frac{w_1}{\sigma_1^2} & \left[ \frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} + \frac{w_3^2}{\sigma_3^2} \right] & -\frac{w_3}{\sigma_3^2} \\ 0 & -\frac{w_3}{\sigma_3^2} & \frac{1}{\sigma_3^2} \end{bmatrix}$$

When we work out the  $2 \times 2$  inverse covariance matrix, all the Blue terms that originated from the child  $y_3$  are lost. So we have

$$\mathbf{K}_2^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & -\frac{w_1}{\sigma_1^2} \\ -\frac{w_1}{\sigma_1^2} & \left[ \frac{1}{\sigma_2^2} + \frac{w_1^2}{\sigma_1^2} \right] \end{bmatrix}$$



Specifically, notice that  $[\mathbf{K}_2^{-1}]_{22}$  is *different* from the (2,2) entry in the three by three  $\mathbf{K}^{-1}$ .

We conclude: *Leaving out a variable leaves  $\mathbf{K}$  unchanged but changes  $\mathbf{K}^{-1}$ .*

This conclusion is important for understanding the answer to the question, ‘*When working with Gaussian processes, why not parameterize the inverse covariance instead of the covariance function?*’

The answer is: you can’t write down the inverse covariance associated with two points! The inverse covariance depends capriciously on *what the other variables are*.

### 3 Energy models

Sometimes people express probabilistic models in terms of energy functions that are minimized in the most probable configuration. For example, in regression with cubic splines, a regularizer is defined which describes the energy that a steel ruler would have if bent into the shape of the curve.

Such models usually have the form:

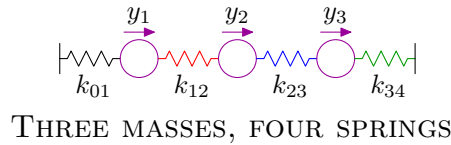
$$P(\mathbf{y}) = \frac{1}{Z} e^{-\frac{E(\mathbf{y})}{T}}, \quad (29)$$

and in simple cases, the energy  $E(\mathbf{y})$  may be a quadratic function of  $\mathbf{y}$ , such as

$$E(\mathbf{y}) = -\sum_{i < j} J_{ij} y_i y_j + \sum_i a_i y_i^2 \quad (30)$$

If so, then the distribution is a Gaussian (just like (1)), and the ‘couplings’  $J_{ij}$  are minus the coefficients in the *inverse* covariance matrix.

As a simple example, consider a set of three masses coupled by springs, and subjected to thermal perturbations.



The equilibrium positions are  $(y_1, y_2, y_3, y_4) = (0, 0, 0, 0)$ , and the spring constants are  $k_{ij}$ . The extension of the second spring is  $y_2 - y_1$ . The energy of this system is

$$\begin{aligned} E(\mathbf{y}) &= \frac{1}{2} k_{01} y_1^2 + \frac{1}{2} k_{12} (y_2 - y_1)^2 + \frac{1}{2} k_{23} (y_3 - y_2)^2 + \frac{1}{2} k_{34} y_3^2 \\ &= \frac{1}{2} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} \begin{bmatrix} k_{01} + k_{12} & -k_{12} & 0 \\ -k_{12} & k_{12} + k_{23} & -k_{23} \\ 0 & -k_{23} & k_{23} + k_{34} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \end{aligned}$$

So at temperature  $T$ , the probability distribution of the displacements is Gaussian with inverse covariance matrix

$$\frac{1}{T} \begin{bmatrix} k_{01} + k_{12} & -k_{12} & 0 \\ -k_{12} & k_{12} + k_{23} & -k_{23} \\ 0 & -k_{23} & k_{23} + k_{34} \end{bmatrix} \quad (31)$$

Notice that there are 0 entries between displacements  $y_1$  and  $y_3$ , the two masses that are not directly coupled by a spring.

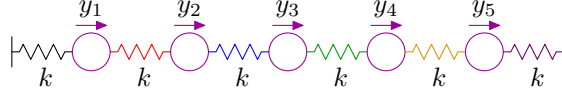


Figure 1. FIVE MASSES, SIX SPRINGS

So inverse covariance matrices are sometimes very sparse. If we have five masses in a row connected by identical springs  $k$  for example, then

$$\mathbf{K}^{-1} = \frac{k}{T} \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}. \quad (32)$$

信息矩阵往往具备这种稀疏结构，而协方差矩阵则不具备

But this sparsity doesn't carry over to the covariance matrix, which is

$$\mathbf{K} = \frac{T}{k} \begin{bmatrix} 0.83 & 0.67 & 0.50 & 0.33 & 0.17 \\ 0.67 & 1.33 & 1.00 & 0.67 & 0.33 \\ 0.50 & 1.00 & 1.50 & 1.00 & 0.50 \\ 0.33 & 0.67 & 1.00 & 1.33 & 0.67 \\ 0.17 & 0.33 & 0.50 & 0.67 & 0.83 \end{bmatrix}. \quad (33)$$

## 4 Eigenvectors and eigenvalues are meaningless

There seems to be a knee-jerk reaction when people see a square matrix: ‘what are its eigenvectors?’ But here, where we are discussing quadratic forms, eigenvectors and eigenvalues have no fundamental status. They are dimensionally invalid objects. Any algorithm that features eigenvectors either didn't need to do so, or shouldn't have done so. (I think the whole idea of principal component analysis is misguided, for example.)

Hang on, you say, what about the three masses example? Don't those three masses have meaningful normal modes? Yes, they do, but those modes are *not* the eigenvectors of the spring matrix (31). Remember, I didn't tell you what the masses of the masses were!

I'm not saying that eigenvectors are never meaningful. What I'm saying is, in the context of quadratic forms

$$\frac{1}{2} \mathbf{y}^T \mathbf{A} \mathbf{y}, \quad (34)$$

eigenvectors are meaningless and arbitrary.

Consider a covariance matrix describing the correlation between something's mass  $y_1$  and its length  $y_2$ .

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix} \quad (35)$$

The dimensions of  $K_{11}$  are mass-squared.  $K_{11}$  might be measured in  $\text{kg}^2$ , for example. The dimensions of  $K_{12} \equiv \langle y_1 y_2 \rangle$  are mass times length.  $K_{12}$  might be measured in  $\text{kg m}$ , for example.

Here's an example, which might describe the correlation between weight and height of some animals in a survey.

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix} = \begin{bmatrix} 10000 \text{ kg}^2 & 70 \text{ kg m} \\ 70 \text{ kg m} & 1 \text{ m}^2 \end{bmatrix} \quad (36)$$

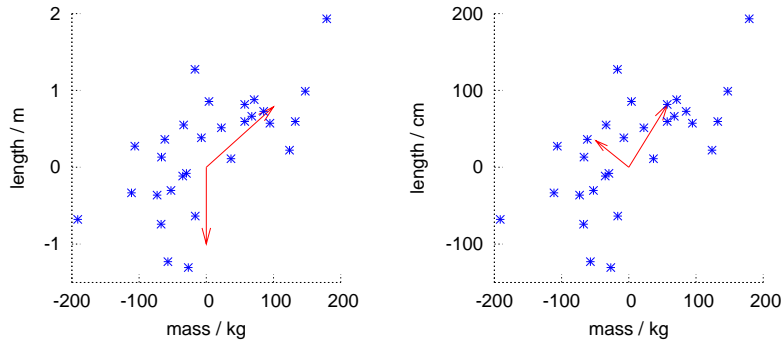


Figure 2. DATASET WITH ITS ‘EIGENVECTORS’. As the text explains, the eigenvectors of covariance matrices are meaningless and arbitrary.

The knee-jerk reaction is “let’s find the principal components of our data”, which means “ignore those silly dimensional units, and just find the eigenvectors of  $\begin{bmatrix} 10000 & 70 \\ 70 & 1 \end{bmatrix}$ ”. But let’s consider what this means. An eigenvector is a vector satisfying

$$\begin{bmatrix} 10000 \text{ kg}^2 & 70 \text{ kg m} \\ 70 \text{ kg m} & 1 \text{ m}^2 \end{bmatrix} \mathbf{e} = \lambda \mathbf{e}. \quad (37)$$

By asking for an eigenvector, we are imagining that two equations are true – first, the top row:

$$10000 \text{ kg}^2 e_1 + 70 \text{ kg m} e_2 = \lambda e_1, \quad (38)$$

and, second, the bottom row:

$$70 \text{ kg m} e_1 + 1 \text{ m}^2 e_2 = \lambda e_2. \quad (39)$$

These expressions violate the rules of dimensions. Try all you like, but you won’t be able to find dimensions for  $e_1$ ,  $e_2$ , and  $\lambda$  such that rule 1 is satisfied.

No, no, the matlab lover says, I leave out the dimensions, and I get:

```
> [e,v] = eig(s)
e =  0.0070002  -0.9999755      v = 5.0998e-01  0.0000e+00
    -0.9999755  -0.0070002      0.0000e+00  1.0000e+04
```

I notice that the eigenvectors  $(0.007, -0.9999)$ , and  $(0.9999, 0.007)$ , which are almost aligned with the coordinate axes. Very interesting! I also notice that the eigenvalues are  $10^4$  and  $0.5$ . What an interestingly large eigenvalue ratio! Wow, that means that there is one very big principal component, and the second one is much smaller. Ooh, how interesting.

But this is nonsense. If we change the units in which we measure length from **m** to cm then the covariance matrix can be written:

$$\mathbf{K} = \begin{bmatrix} K_{11} & K_{12} \\ K_{12} & K_{22} \end{bmatrix} = \begin{bmatrix} 10000 \text{ kg}^2 & 7000 \text{ kg cm} \\ 7000 \text{ kg cm} & 10000 \text{ cm}^2 \end{bmatrix} \quad (40)$$

This is exactly the same covariance matrix of exactly the same data. But the eigenvectors and eigenvalues are now:

$$\begin{array}{cc} \mathbf{e} = & \begin{matrix} -0.70711 & 0.70711 \\ 0.70711 & 0.70711 \end{matrix} & \mathbf{v} = & \begin{matrix} 3000 & 0 \\ 0 & 17000 \end{matrix} \end{array}$$

Figure 2 illustrates this situation.

On the left, a data set of masses and lengths measured in metres. The arrows show the ‘eigenvectors’. (The arrows don’t look ‘orthogonal’ in this plot because a step of one unit on the x-axis happens to cover less paper than a step of one unit on the y-axis.) On the right, exactly the same data set but with lengths measured in centimetres. The arrows show the ‘eigenvectors’.

In conclusion, eigenvectors of the matrix in a quadratic form are not fundamentally meaningful. [Properties of that matrix that *are* meaningful include its determinant.]

## 4.1 Aside

This complaint about eigenvectors comes hand in hand with another complaint, about ‘steepest descent’. A steepest descent algorithm is dimensionally invalid. A *step* in a parameter space does not have the same dimensions as a *gradient*. To turn a gradient into a sensible step direction, you need a *metric*. The metric defines how ‘big’ a step is (in rather the same way that when gnuplot plotted the data above, it chose a vertical scale and a horizontal scale). Once you know how big alternative steps are, it becomes meaningful to take the step that is ‘steepest’ (that is, it’s the direction with the biggest change in function value per unit ‘distance’ moved).

Without a metric, steepest descents algorithms are not *covariant*. That is, the algorithm would behave differently if you just changed the units in which one parameter is measured.

## Appendix: Answers to quiz

For the first four, you can quickly guess the answers based on whether the (1,3) entries are zero or not. For a careful answer you should also check that the matrices really are positive definite (they are) and that they are realisable by the respective graphical models (which isn’t guaranteed by the preceding constraints).

1. A and B
2. C and D
3. C and D
4. A and B
5. A is true, B is false.